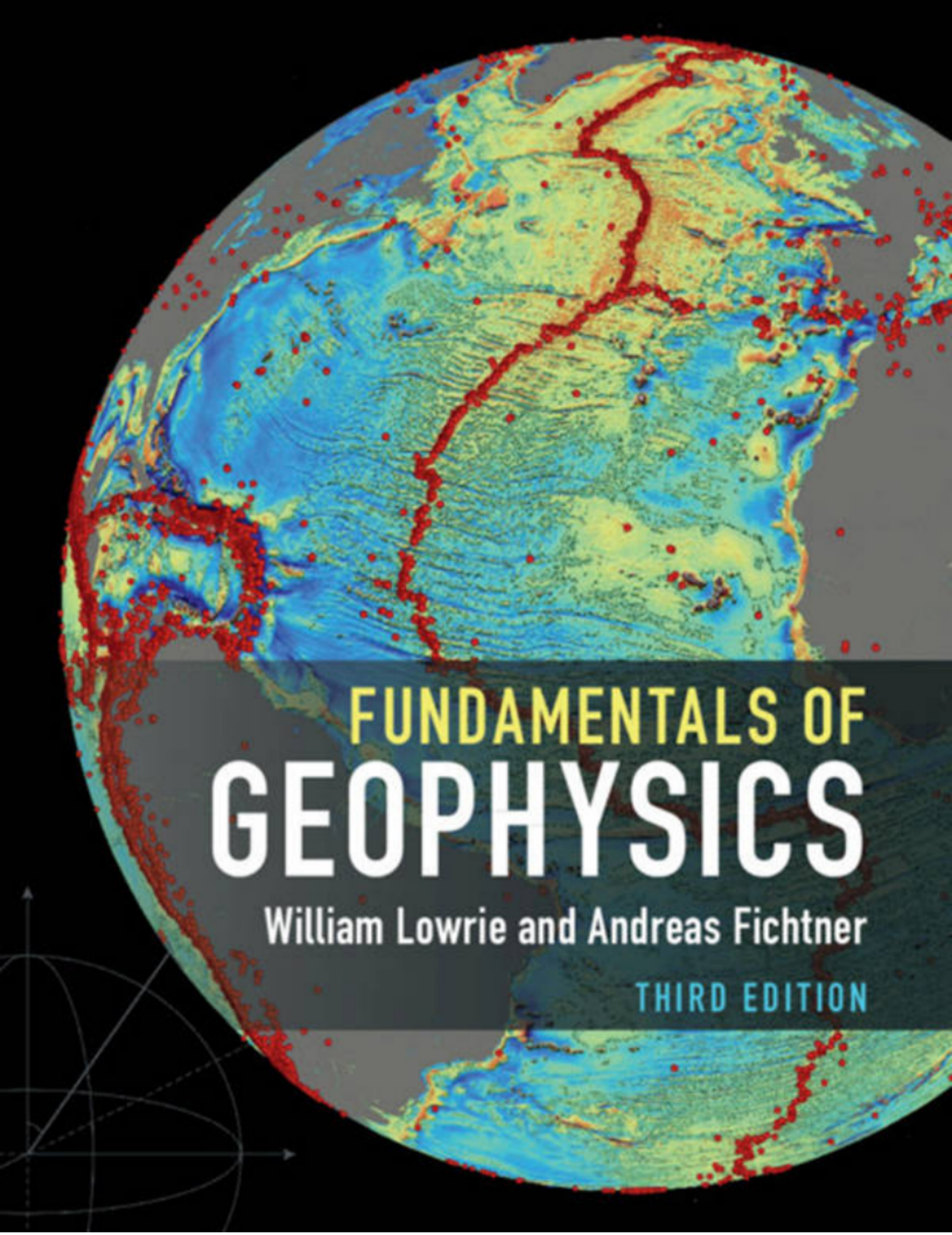


FUNDAMENTALS OF GEOPHYSICS

William Lowrie and Andreas Fichtner

THIRD EDITION

Fundamentals of Geophysics

Third Edition

This enduringly popular undergraduate textbook has been thoroughly reworked and updated, and now comprises twelve chapters covering the same breadth of topics as earlier editions, but in a substantially modernized fashion to facilitate classroom teaching.

Covering both theoretical and applied aspects of geophysics, clear explanations of the physical principles are blended with step-by-step derivations of the key equations and over 400 explanatory figures to explain the internal structure and properties of the planet, including its petroleum and mineral resources. New topics include the latest data acquisition technologies, such as satellite geophysics, planetary landers, ocean-bottom seismometers, and fiber optic methods, as well as recent research developments in ambient noise interferometry, seismic hazard analysis, rheology, and numerical modeling – all illustrated with examples from the scientific literature.

Student-friendly features include separate text boxes with auxiliary explanations and advanced topics of interest; reading lists of foundational, alternative, or more detailed resources; end-of-chapter review questions; and an increased number of quantitative exercises. Completely new to this edition is the addition of computational exercises in Python, designed to help students acquire important programming skills and develop a more profound understanding of geophysics.

William Lowrie is Professor Emeritus of Geophysics at the Institute of Geophysics at ETH Zürich, Switzerland, where he taught geophysics and carried out research in rock magnetism and paleomagnetism. He graduated from the University of Edinburgh in 1960 with first class honors in physics, before completing a master's degree in geophysics at the University of Toronto and a doctorate at the University of Pittsburgh. He has been President of the European Union of

Geosciences (1987–1989) and Section President and a council member of the American Geophysical Union (2000–2002). He is a fellow of AGU and a member of the Academia Europaea.

Andreas Fichtner is Professor of Geophysics at ETH Zürich. He received his PhD in 2010 from the University of Munich. His principal research interests include the development and application of methods for full seismic waveform inversion, resolution analysis in tomography, earthquake source inversion, seismic interferometry, and inverse theory. For his work, he received the Keiiti Aki Award from the American Geophysical Union, the Early Career Scientist Award from the International Union of Geodesy and Geophysics, and the Hoffmann Prize from the Bavarian Academy of Sciences. He is a former Fulbright scholar and a current member of the Young Academy of Europe.

Fundamentals of Geophysics

Third Edition

William Lowrie
Swiss Federal Institute of Technology Zürich

Andreas Fichtner
Swiss Federal Institute of Technology Zürich

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781108492737

DOI: 10.1017/9781108685917

First edition © Cambridge University Press 1997

Second edition © W. Lowrie 2007

Third edition © Cambridge University Press 2020

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 1997

Second edition 2007

9th printing 2017

Third edition 2020

Printed in the United Kingdom by TJ International Ltd, Padstow Cornwall 2020

A catalogue record for this publication is available from the British Library.

ISBN 978-1-108-49273-7 Hardback

ISBN 978-1-108-71697-0 Paperback

Additional resources for this publication at www.cambridge.org/FoG3e.

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

This book is dedicated to Marcia and Carolin

Contents

Preface

1The Solar System

Preview

1.1The Planets

1.2The Discovery of the Planets and Determination of Their Orbits

1.3Kepler's Laws of Planetary Motion

1.4Characteristics of the Planets and Their Orbits

1.5The Inner (Terrestrial) Planets and the Moon

1.6The Outer (Giant) Planets

1.7The Outer Solar System: Trans-Neptunian Objects

1.8Suggestions for Further Reading

1.9Review Questions

1.10Exercises

2Plate Tectonics

Preview

2.1Historical Introduction

2.2Continental Drift

2.3Earth Structure

2.4Types of Plate Margin

2.5Sea-Floor Spreading

2.6Plate Margin Features

2.7Triple Junctions

2.8Hotspots

2.9Plate Motion on the Surface of a Sphere

2.10Forces Driving Plate Tectonic Motions

2.11Suggestions for Further Reading

- 2.12 Review Questions
- 2.13 Exercises
- 2.14 Computational Exercises

3 Gravity and the Figure of the Earth
- Preview
- 3.1 The Earth’s Size and Shape
- 3.2 Gravitation
- 3.3 The Earth’s Rotation
- 3.4 The Earth’s Figure and Gravity
- 3.5 Space Geodesy
- 3.6 Suggestions for Further Reading
- 3.7 Review Questions
- 3.8 Exercises
- 3.9 Computational Exercises

4 Gravity Surveying
- Preview
- 4.1 Gravity Measurement and Reduction
- 4.2 Interpretation of Gravity Anomalies
- 4.3 Isostasy
- 4.4 Suggestions for Further Reading
- 4.5 Review Questions
- 4.6 Exercises
- 4.7 Computational Exercises

5 Rheology of the Earth
- Preview
- 5.1 Elastic Deformation
- 5.2 Viscous Flow
- 5.3 Deviations from Perfect Elasticity and Viscous Flow
- 5.4 Lithosphere Rigidity
- 5.5 Mantle Viscosity
- 5.6 Suggestions for Further Reading

- 5.7 Review Questions
- 5.8 Exercises
- 5.9 Computational Exercises

6 Seismology
- Preview
- 6.1 Introduction
- 6.2 Seismic Waves
- 6.3 Seismic Waves in the Heterogeneous Earth
- 6.4 Ambient Seismic Waves
- 6.5 Seismometry
- 6.6 Suggestions for Further Reading
- 6.7 Review Questions
- 6.8 Exercises
- 6.9 Computational Exercises

7 Earthquakes and the Earth’s Internal Structure
- Preview
- 7.1 Introduction
- 7.2 Earthquake Seismology
- 7.3 Earthquakes and Human Society
- 7.4 Internal Structure of the Earth
- 7.5 Suggestions for Further Reading
- 7.6 Review Questions
- 7.7 Exercises
- 7.8 Computational Exercises

8 Geochronology
- Preview
- 8.1 Time
- 8.2 Historical Estimates of the Earth’s Age
- 8.3 Radioactivity
- 8.4 Radiometric Age Determination
- 8.5 Ages of the Earth and the Solar System

8.6Suggestions for Further Reading
8.7Review Questions
8.8Exercises
8.9Computational Exercises

9The Earth’s Heat
Preview
9.1Introduction
9.2Thermodynamic Principles
9.3Temperature Inside the Earth
9.4Heat Transport in the Earth
9.5Sources of Heat in the Earth
9.6The Heat Conduction Equation
9.7Continental Heat Flow
9.8Oceanic Heat Flow
9.9Mantle Convection
9.10Suggestions for Further Reading
9.11Review Questions
9.12Exercises
9.13Computational Exercises

10Geoelectricity
Preview
10.1Introduction
10.2Electrical Principles
10.3Electrical Properties of the Earth
10.4Natural Potentials and Currents
10.5Resistivity Surveying
10.6Electromagnetic Surveying
10.7Electrical Conductivity in the Earth
10.8Suggestions for Further Reading
10.9Review Questions
10.10Exercises
10.11Computational Exercises

11 The Earth’s Magnetic Field

Preview

- **11.1** Magnetism
- **11.2** Geomagnetism
- **11.3** Magnetic Fields of the Sun, Moon, and Planets
- **11.4** Magnetic Surveying
- **11.5** Suggestions for Further Reading
- **11.6** Review Questions
- **11.7** Exercises
- **11.8** Computational Exercises

12 Paleomagnetism

Preview

- **12.1** Rock Magnetism
- **12.2** Apparent Polar Wander and Tectonics
- **12.3** Geomagnetic Polarity
- **12.4** Suggestions for Further Reading
- **12.5** Review Questions
- **12.6** Exercises
- **12.7** Computational Exercises

Appendices

- ***A*** *The Three-Dimensional Wave Equation*
- ***B*** *Cooling of a Semi-Infinite Half-Space*
- ***C*** *Magnetic Behavior of Rock-Forming Minerals*
- ***D*** *Magnetic Anisotropy*

Bibliography

Index

Preface

Widespread access to powerful computers, advances in instrumentation, and the expansion of remote sensing of properties of the Earth and some other planets from space missions have resulted in considerable advancements in many fields of geophysics in the years since the second edition of this book was published. During this time the continual spread of internet-related education has made online availability of teaching materials increasingly desirable. These developments encouraged us to prepare this third edition of *Fundamentals of Geophysics*, and to provide interactive involvement for students in solving problems online in the form of Jupyter Notebooks.

For this edition, William Lowrie, the author of the earlier editions, has been joined by Andreas Fichtner. Both authors are professors of geophysics at the Swiss Federal Institute of Technology (ETH Zürich), with different fields of specialization. The collaboration adds expertise and a fresher approach to the different disciplines covered in the book.

Geophysics is often taught as two topics, with titles like “general” and “applied.” In fact these are two sides of the same coin, and they often benefit from advances in instrumentation and analysis developed in each other’s domain. As in the earlier editions, we therefore continue to describe the principal methods in both fields.

It is not possible to explain briefly every recent advancement in an introductory textbook, but we have included some of the most striking. These include the geophysical results from space missions to other bodies in the solar system (e.g., Mercury, Mars, Jupiter, and Pluto). There has been great progress in remotely sensing the gravitational and magnetic fields of our own planet from orbiting satellites. A new section is added to deal with space geodesy, in which we explain some of the developments that have revolutionized research in geodesy and gravity. New seismic methods for understanding the Earth's internal structure are described, for example by analyzing ambient noise. The danger presented by earthquakes can be estimated by probabilistic seismic hazard. Advancements in seismic tomography and the numerical modeling of geophysical processes (e.g., mantle dynamics) have illuminated – and posed new questions about – our understanding of the Earth's interior. Text boxes are used to handle some topics in greater detail than is needed in the body of the text, which, as in the earlier editions, emphasizes fundamental principles in the individual disciplines.

The new structure of the book divides the field of geophysics into 12 chapters, with a little overlap between some of them. In order to keep the price to students as low as possible, the authors decided to continue to use black-and-white figures, avoiding the use of more expensive color. However, as the book's title claims, its emphasis is on teaching the fundamental principles and these (mostly) do not change. If necessary, the teacher is in a position to enlarge upon the illustrations with colored examples, of which there is abundant choice in the professional literature. The website provides the solutions to the exercises at the end of each chapter for teachers who want to use them.

In preparing this edition we have received contributed figures, voluntary reviews, and constructive suggestions from a large number of colleagues. We are very grateful for their generous support. If we have not accepted suggestions, it was usually not because we disagreed, but rather because of our personal preferences. In particular we wish to thank Michael Afanasiev, Peter Annan, Nienke Blom, Jim Channell, Rob Coe, Laurentiu Danciu, Rhodri Davies, Sjoerd de Ridder, Jordi Diaz, Laura Ermert, Chris Finlay, Alexandre Fournier, Domenico Giardini, Alan Green, James Harris, Ann Hirt, Ian Jackson, Dennis Kent, Paula Koelemeijer, Maria Koroni, Kostas Lentas, Guust Nolet, Markku Poutanen, and Andrew Schaeffer.

In addition, the book has benefited from the suggestions of anonymous reviewers of each chapter. We appreciate the time and effort they made to improve our book, surely at the expense of their involvement in other research or academic work, and thank them sincerely for their help.

Finally, we thank our wives, Marcia Lowrie and Carolin Fichtner, for their understanding and encouragement. This book is dedicated to them.

1

The Solar System

Preview

The solar system originated more than 4.5 billion years ago from a rotating cloud of molecular hydrogen and interstellar dust surrounding a star that became the Sun. Collisions and coalescence of particles in the dust cloud led to the formation of a solar system with eight planets and countless smaller bodies that has fascinated mankind since prehistoric times. Our knowledge of the solar system exploded in the recent era of space exploration, and its complete description deserves a library on its own. This chapter provides brief descriptions of the orbits and physical properties of the major planetary bodies. These exhibit natural phenomena originally studied on the Earth, such as gravitational and magnetic fields, seismicity, and tidal motions. They can be understood with geophysical concepts and techniques introduced in later chapters.

1.1 The Planets

The solar system consists of the Sun, the planets and their moons, and other objects that are in orbit around the Sun under its gravitational attraction. The eight planets form two groups, differing in size, composition, and proximity to the Sun (Table 1.1). The four inner (or terrestrial) planets – Mercury, Venus, Earth, and Mars – resemble the Earth in size and density (Fig. 1.1). They have a solid, rocky composition and they rotate about their own axes at the same rate or slower than the Earth. The four outer (or giant) planets – Jupiter, Saturn, Uranus, and Neptune – are much larger than the Earth, with lower densities and higher rotation rates. The inner planets have few moons and no rings; in contrast, the outer planets have many moons and are surrounded by rings of dust. Between Mars and Jupiter lies the asteroid belt, made up of numerous orbiting objects, some as small as dust particles and others up to hundreds of kilometers in diameter.

Table 1.1 Dimensions and rotational characteristics of the planets and the Moon

The great planets and the dwarf planet Pluto are gaseous. For these planets the surface on which the pressure is 1 atmosphere is taken as the effective radius. In the definition of polar flattening, a and c are respectively the semi-major and semi-minor axis of the spheroidal shape.

Planet	**Mass M (10^{24} kg)**	**Mass relative to the Earth**	**Mean density (kg m−^{3})**	**Equatorial radius (km)**	**Sidereal rotation period (days)**	**Inverse flattening $f^{-1} = a/(a-c)$**	**Obliquity of axis to orbit (°)**
Terrestrial planets and the Moon							
Mercury	0.33011	0.0553	5427	2439.7	58.65	–	0.034

Venus	4.8675	0.815	5243	6051.8	243.7	–	177.4
Earth	5.9723	1.000	5514	6378.1	0.9973	298.253	23.44
Moon	0.07346	0.0123	3344	1738.1	27.32	827.67	6.68
Mars	0.64171	0.1074	3933	3396.2	1.026	169.81	25.19
Great planets and Pluto							
Jupiter	1898.19	317.83	1326	71,492	0.414	15.41	3.13
Saturn	568.34	95.16	687	60,268	0.444	10.21	26.73
Uranus	86.813	14.54	1271	25,559	0.718	43.62	97.77
Neptune	102.413	17.15	1638	24,764	0.671	58.54	28.32
Pluto	0.01303	0.00218	1854	1187	6.387	–	122.53

Source: NASA Space Science Data Center, 2017, https://nssdc.gsfc.nasa.gov/planetary.

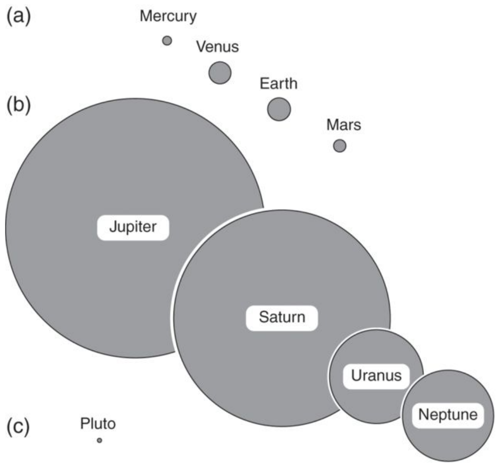


Fig. 1.1 The relative sizes of the planets: (a) the four terrestrial planets, (b) the four giant planets, and (c) the dwarf planet Pluto, which is diminutive compared to the others.

In addition to the eight *major planets* and Pluto, there are other large objects in orbit around the Sun. Designated as *minor planets*, they do not fulfill the criteria common to the definition of the major planets. The discovery of large objects in the solar system beyond the orbit of Neptune stimulated astronomers to debate the criteria that should be used to define a planet. In 2006 the International

Astronomical Union decided that to be a planet in the solar system an object must satisfy three conditions: (1) it must be in orbit around the Sun; (2) it must be large enough that its own gravitation results in a spherical or spheroidal shape; and (3) it must have cleared the neighborhood around its orbit of planetesimals. An object that meets conditions (1) and (2) and is not a satellite of another body is classified as a “dwarf planet.” Pluto’s elongate orbit is affected by Neptune and the Kuiper belt bodies, so it fails condition (3) and is now classified as a dwarf planet. Other dwarf planets are the largest asteroid, Ceres, which has a radius of 950 km, and the scattered disk object Eris (Section 1.5.6).

The orbits of the planets around the Sun lie close to each other, forming a flat disk. The plane of the Earth’s orbit is called the *ecliptic plane*; the orbits of the other planets lie within a few degrees of it. The shapes of the orbits are slightly elliptical, deviating by only a small amount from being circles. The shapes of the orbits and the motions of the planets are controlled by the laws of conservation of energy and angular momentum, and by the inverse square law of gravitation.

1.1.1 Conservation of Energy and Angular Momentum

To a first approximation the Sun and planets form an isolated system. This is one in which no matter or energy is transferred into or out of the system. The law of *conservation of energy* states that the total energy of an isolated system is constant. In the solar system the planets are bound to the Sun by its gravitation, which varies as the inverse square of distance from the Sun's center. These conditions result in each planet having a planar orbit with an elliptical shape (Box 1.1).

A further constraint on planetary motions is the law of *conservation of angular momentum*. This applies to rotating systems and states that the angular momentum of a system remains constant unless it is acted on by an external torque (or turning force). Just as the linear momentum of a body depends on its velocity and mass, the angular momentum of a rotating body depends on its rate of rotation about an axis and the distribution of its mass about that axis, a property known as moment of inertia.

For a point mass m the moment of inertia (I) about an axis at distance r is defined as:

$$I = mr^2. \tag{1.1}$$

The angular momentum (h) is defined as the product of the moment of inertia (I) about an axis and the rate of rotation (Ω) about that axis:

$$h = I\Omega . \tag{1.2}$$

Each planet revolves in a nearly circular orbit around the Sun and at the same time rotates about its own axis. Thus there are two contributions to its angular momentum. The angular momentum of a planet's revolution about the Sun is obtained quite simply. The solar system is so immense that the physical size of each planet is tiny compared to the size of its orbit. The moment of inertia of a planet about the Sun is computed by inserting the mass of the planet and its orbital radius (Table 1.2) in Eq. (1.1); the orbital angular momentum of the planet follows by combining the computed moment of inertia with the rate of orbital revolution, as in Eq. (1.2).

Table 1.2 Dimensions and characteristics of the planetary orbits

1 AU = 1 astronomical unit = 149,597,870.7 km.

Planet	Mean orbital radius (AU)	Semi-major axis (10^6 km)	Eccentricity of orbit	Inclination of orbit to ecliptic (°)	Mean orbital velocity (km s^{-1})	Sidereal period of orbit (yr)
Terrestrial planets and the Moon						
Mercury	0.3830	57.91	0.2056	7.00	47.36	0.2408
Venus	0.7233	108.21	0.00677	3.39	35.02	0.6152
Earth	1.0000	149.60	0.01671	0.000	29.78	1.0000
Moon (about Earth)	0.00257	0.3844	0.0549	5.145	1.022	0.0748
Mars	1.5202	227.92	0.0935	1.851	24.08	1.8808
Great planets and Pluto						
Jupiter	5.2013	778.6	0.0489	1.304	13.06	11.862
Saturn	9.575	1433.5	0.0565	2.485	9.68	29.457
Uranus	19.19	2872.6	0.0457	0.772	6.80	84.011
Neptune	30.05	4495.1	0.0113	1.769	5.43	164.79
Pluto	38.86	5906.4	0.2488	17.16	4.67	247.94

Source: National Space Science Data Center, 2017, https://nssdc.gsfc.nasa.gov/planetary.

To determine the moment of inertia of a solid body about an axis that passes through it (e.g., the rotational axis of a planet) is more complicated. Equation (1.2) must be computed and summed for all particles in the planet. If the planet is represented by a sphere of mass M and mean radius R, the moment of inertia C about the axis of rotation is given by

$$C = kMr^2, \tag{1.3}$$

where the constant k is determined by the density distribution within the planet. Its value can be calculated exactly for simple geometrical shapes. For example, if the density is uniform inside a solid sphere, the value of k is exactly 2/5, or 0.4; for a hollow sphere it is 2/3. If density increases with depth in a planet, e.g., if it has a dense core, the value of k is less than 0.4; for the Earth, $k = 0.3308$. For some planets the variation of density with depth is not well known, but for most planets there is enough information to calculate the moment of inertia about the axis of rotation; combined with the rate of rotation as in Eq. (1.2), this gives the rotational angular momentum (Table 1.3).

Table 1.3 Distributions of orbital and rotational angular momentum in the solar system

Note that Venus, Uranus, and Pluto have retrograde axial rotations.

	Planet mass M (10^{24} kg)	**Mean orbital rate ω (10^{-9} rad s^{-1})**	**Mean orbital radius r (10^{9} m)**	**Orbital angular momentum $M\omega r^2$ (10^{39} kg m^2 s^{-1})**	**Planet mean radius R (10^{6} m)**	**Normalized moment of inertia I/MR^2**	**Moment of inertia I (10^{40} kg m^2)**	**Axial rotation rate Ω (10^{-6} rad s^{-1})**	**Rotational angular momentum $I\Omega$ (10^{39} kg m^2 s^{-1})**
Terrestrial planets									
Mercury	0.3301	827.3	57.29	0.896	2.440	0.353	6.88×10^{-5}	1.240	8.53×10^{-10}
Venus	4.8675	323.9	108.2	18.46	6.052	0.33	5.88	–	1.76

							$\times 10^{-3}$	0.2992	$\times 10^{-8}$
Earth	5.9723	199.2	149.6	26.63	6.371	0.3308	8.02×10^{-3}	72.92	5.85×10^{-6}
Mars	0.6417	105.9	227.4	3.52	3.390	0.366	2.70×10^{-4}	70.78	1.91×10^{-7}
Great planets and Pluto									
Jupiter	1898.2	16.8	778.1	19,304	69.911	0.254	235.6	175.7	0.4144
Saturn	568.3	6.76	1432	7887	58.232	0.21	40.5	163.8	0.0663
Uranus	86.8	2.37	2871	1697	25.362	0.23	1.28	– 101.3	0.00124
Neptune	102.4	1.209	4495	2502	24.622	0.24	1.49	108.1	0.00167
Pluto	0.01303	0.804	5814	0.354	1.187	–	–	–11.4	–
TOTAL	2668	–	–	31,439	–	–	–	–	0.483
The Sun	1,989,100	–	–	–	695.700	0.070	6,737,000	2.865	162.6

Source: National Space Science Data Center, 2017, http://nssdc.gsfc.nasa.gov/planetary.

The angular momentum of a planet's revolution about the Sun is much greater (on average about 60,000 times) than the angular momentum of its rotation about its own axis. Whereas more than 99.9% of the total mass of the solar system is concentrated in the Sun, more than 99% of the angular momentum is carried by the orbital motion of the planets, especially the four giant planets. Of these, Jupiter is a special case: it accounts for over 70% of the mass and more than 60% of the angular momentum of the planets.

1.1.2 Origin of the Solar System

Numerous theories have been proposed for the origin of the solar system. Age determinations on meteorites indicate that the solar system originated about 4.5–4.6 × 10^9 years ago. A successful theory of how it originated must account satisfactorily for the observed characteristics of the planets. The most important of these properties are:

1. The planetary orbits lie in or close to the same plane, which contains the Sun and the orbit of the Earth (the ecliptic plane).
2. The planets revolve about the Sun in the same sense, which is counterclockwise when viewed from above the ecliptic plane. This sense of rotation is defined as prograde.
3. The rotations of the planets about their own axes are also mostly prograde. The exceptions are Venus, which has a retrograde rotation; and Uranus, whose axis of rotation lies nearly in the plane of its orbit.
4. Each planet is roughly twice as far from the Sun as its closest neighbor.
5. The compositions of the planets make up two distinct groups: the terrestrial planets lying close to the Sun are small and have high densities, whereas the giant planets far from the Sun are large and have low densities.
6. The Sun has almost 99.9% of the mass of the solar system, but the planets account for more than 99% of the angular momentum.

The first theory based on scientific observation was the *nebular hypothesis* introduced by the German philosopher Immanuel Kant in 1755 and formulated by the French astronomer Pierre Simon de Laplace in 1796. According to this hypothesis the planets and their satellites were formed at the same time as the Sun. Space was filled by a rotating cloud (*nebula*) of hot primordial gas and dust that, as it cooled, began to contract. To conserve the angular momentum of the system, its rotation speeded up; a familiar analogy is the way a pirouetting skater spins more rapidly when he draws in his outstretched arms. Centrifugal force would have caused concentric rings of matter to be thrown off, which then condensed into planets. A serious objection to this hypothesis is that the mass of material in each ring would be too small to provide the gravitational attraction needed to cause the ring to condense into a planet. Moreover, as the nebula contracted, the largest part of the angular momentum would remain associated with the main mass that condensed to form the Sun, which disagrees with the observed distribution of angular momentum in the solar system.

Several alternative models were postulated subsequently, but have also fallen into disfavor. For example, the *collision hypothesis* assumed that the Sun was formed before the planets. The gravitational attraction of a closely passing star or the blast of a nearby supernova explosion drew out a filament of solar material that condensed to form the planets. However, a major objection to this scenario is that the solar material would have been so hot that it dissipated explosively into space rather than slowly condensing to form the planets.

Current interpretations of the origin of the solar system are based on modifications of the nebular hypothesis. As the cloud of gas and dust contracted, its rate of rotation speeded up, flattening the cloud into a lens-shaped disk. When the core of the contracting cloud became dense enough, gravitation caused it to collapse upon itself to form a proto-Sun in which thermonuclear fusion was initiated. Hydrogen nuclei combined under the intense pressure to form helium nuclei, releasing huge amounts of energy. The material in the spinning disk was initially very hot and gaseous but, as it cooled, solid material condensed out of it as small grains. The grains coalesced as rocky or icy clumps called *planetesimals*. Asteroid-like planetesimals with a silicate, or rocky, composition formed near the Sun, while comet-like planetesimals with an icy composition formed far from the Sun’s heat. Collisions and the gravitational attraction between the planetesimals resulted in their further accretion to form the planets. Matter with a high boiling point (e.g., metals and silicates) could condense near to the Sun, forming the terrestrial planets.

The terrestrial planets acquired an internal structural of concentric spherical shells, determined by the chemical composition and density of their constituents. Gravitational forces caused denser materials to sink toward the center of the planet, forming an iron-rich core. Less-dense silicates formed a surrounding shell (or mantle) around the core, with a thin solid outer crust of silicates. Above this layered structure a liquid hydrosphere accumulated, and above it a gaseous atmosphere. The Earth’s atmosphere is itself layered, consisting of the troposphere (responsible for weather), stratosphere (containing the protective ozone layer), and the ionosphere (formed by ionization of air molecules by solar radiation). The other planets and some of their moons also have atmospheres with differing compositions.

By contrast to the terrestrial planets, volatile materials (e.g., water, ammonia, methane) were vaporized and driven into space by the stream of particles and radiation from the Sun. Between Mars and Jupiter, at a distance known as the frost line, the warming effect of the solar radiation becomes so low that volatile compounds are solid, in the form of ices. During the condensation of the large cold planets in the frigid distant realms of the solar system, the volatile materials were retained. The gravitational attractions of Jupiter and Saturn may have been strong enough to retain the composition of the original nebula.

It is important to keep in mind that this scenario is a hypothesis – a plausible but not unique explanation of how the solar system formed. It attributes the variable compositions of the planets to accretion at different distances from the Sun. The model can be embellished in many details to account for the characteristics of individual planets. However, the scenario is unsatisfactory because it is mostly qualitative. For example, it does not explain adequately the distribution of angular momentum. Physicists, astronomers, mathematicians, and space scientists are constantly trying new methods of investigation and searching for additional clues that will improve the hypothesis of how the solar system formed.

1.2 The Discovery of the Planets and Determination of Their Orbits

To appreciate how impressive the night sky must have been to early man it is necessary today to go to a place remote from the distracting lights and pollution of urban centers. Viewed from the wilderness, the firmaments appear to the naked eye as a canopy of shining points, fixed in space relative to each other. Early observers noted that the star pattern appeared to move regularly and used this as a basis for determining the timing of events. More than 3000 years ago, in about the thirteenth century BC, the year and month were combined in a working calendar by the Chinese, and about 350 BC the Chinese astronomer Shih Shen prepared a catalog of the positions of 800 stars. The ancient Greeks observed that several celestial bodies moved back and forth against this fixed background and called them the *planetes*, meaning “wanderers.” In addition to the Sun and Moon, the naked eye could discern the planets Mercury, Venus, Mars, Jupiter, and Saturn.

Geometrical ideas were introduced into astronomy by the Greek philosopher Thales in the sixth century BC. This advance enabled the Greeks to develop astronomy to its highest point in the ancient world. Aristotle (384–322 BC) summarized the Greek work performed prior to his time and proposed a model of the universe with the Earth at its center. This *geocentric* model became embedded in religious conviction and remained in authority until late into the Middle Ages. It did not go undisputed; Aristarchus of Samos (*c.*310–*c.*230 BC) determined the sizes and distances of the Sun and Moon relative to the Earth and proposed a *heliocentric* (sun-centered) cosmology. The methods of trigonometry developed by Hipparchus (190–120 BC) enabled the determination of astronomical distances by observing the angular positions of celestial bodies. Ptolemy, a Greco-Egyptian astronomer in the second century AD, applied these methods to the known planets and was able to predict their motions with remarkable accuracy, considering the primitiveness of available instrumentation.

Until the invention of the telescope in the early seventeenth century the main instrument used by astronomers for determining the positions and distances of heavenly bodies was the *astrolabe*. This device consisted of a disk of wood or metal with the circumference marked off in degrees. At its center was pivoted a movable pointer called the *alidade*. Angular distances could be determined by sighting on a body with the alidade and reading off its elevation from the graduated scale. The inventor of the astrolabe is not known, but it is often ascribed to Hipparchus (190–120 BC). It remained an important tool for navigators until the invention of the sextant in the eighteenth century.

The angular observations were converted into distances by applying the method of parallax. This is simply illustrated by the following example. Consider the planet P as viewed from the Earth at different positions in the latter's orbit around the Sun (Fig. 1.2). For simplicity, treat planet P as a stationary object (i.e., disregard the planet's orbital motion). The angle between a sighting on the planet and on a fixed star will appear to change because of the Earth's orbital motion around the Sun. Let the measured extreme angles be θ_1 and θ_2 and the distance of the Earth from the Sun be s; the distance between the extreme positions E and E′ of the orbit is then $2s$. The distances p_1 and p_2 of the planet from the Earth are computed in terms of the Earth–Sun distance by applying the trigonometric law of sines:

$$
\begin{aligned}
\frac{p_1}{2s} &= \frac{\sin\left(\frac{\pi}{2}-\theta_2\right)}{\sin(\theta_1+\theta_2)} = \frac{\cos\theta_2}{\sin(\theta_1+\theta_2)} \\
\frac{p_2}{2s} &= \frac{\cos\theta_1}{\sin(\theta_1+\theta_2)}.
\end{aligned}
\tag{1.4}
$$

Further trigonometric calculations give the distances of the planets from the Sun. The principle of parallax was also used to determine relative distances in the Aristotelian geocentric system, according to which the fixed stars, Sun, Moon, and planets are considered to be in motion about the Earth.

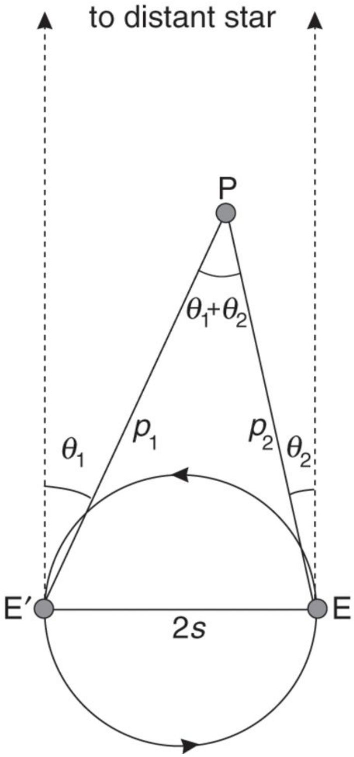

to distant star
P
$\theta_1+\theta_2$
θ_1
p_1
p_2
θ_2
E′
E
2s

Fig. 1.2 The method of parallax, in which two measured angles (θ_1 and θ_2) are used to compute the distances (p_1 and p_2) of a planet from the Earth in terms of the Earth–Sun distance (s).

1.2.1 Bode's Law

In 1772 the German astronomer Johann Bode devised an empirical formula to express the approximate distances of the planets from the Sun. A series of numbers is created in the following way: the first number is zero, the second is 0.3, and the rest are obtained by doubling the previous number. This gives the sequence 0, 0.3, 0.6, 1.2, 2.4, 4.8, 9.6, 19.2, 38.4, 76.8. Each number is then augmented by 0.4 to give the sequence: 0.4, 0.7, 1.0, 1.6, 2.8, 5.2, 10.0, 19.6, 38.8, 77.2. This series can be expressed mathematically as follows:

$$\begin{aligned} d_n &= 0.4 \quad \text{for} \quad n = 1 \\ d_n &= 0.4 + 0.3 \times 2^{n-2} \quad \text{for} \quad n \geq 2\,. \end{aligned} \tag{1.5}$$

This expression gives the distance d_n in astronomical units (AU) of the nth planet from the Sun. It is usually known as Bode's law but, as the same relationship was suggested earlier by J. D. Titius of Wittenberg, it is sometimes called Titius–Bode's law. Examination of Fig. 1.3 and comparison with Table 1.2 shows that this relationship holds remarkably well, except for Neptune and Pluto, the latter of which has been demoted to the status of a dwarf planet. It has been suggested that the orbits of these planets are no longer their original orbits.

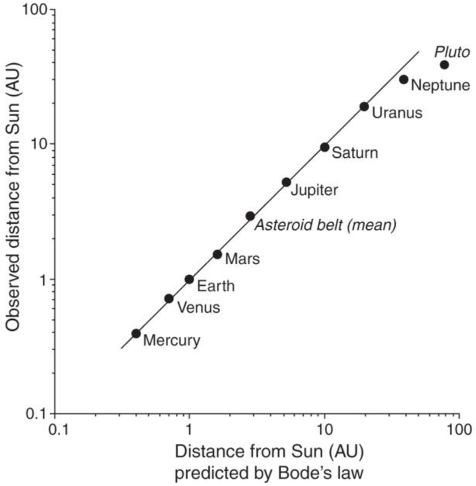


Fig. 1.3 The empirical relationship of Titius–Bode for the distances of the planets from the Sun.

Bode's law predicts a fifth planet at 2.8 AU from the Sun, between the orbits of Mars and Jupiter. In the last years of the eighteenth century astronomers searched intensively for this missing planet. In 1801 a small planetoid, Ceres, was found at a distance of 2.77 AU from the Sun. Subsequently, it was found that numerous small planetoids occupied a broad band of solar orbits centered around 2.9 AU, now called the *asteroid belt*. Pallas was found in 1802, Juno in 1804, and Vesta, the only asteroid that can be seen with the naked eye, was found in 1807. By 1890 more than 300 asteroids had been identified. In 1891 astronomers began to record their paths on photographic plates. Thousands of asteroids occupying a broad belt between Mars and Jupiter, at distances of 2.15–3.3 AU from the Sun, have since been tracked and cataloged.

Bode's "law" is not a law in the scientific sense, and is therefore often referred to as Bode's rule. Though a simple explanation of Bode's law currently does not exist, recent investigations indicate that the complicated gravitational interaction of many planets may indeed lead to an orbital spacing that approximately follows a power law. During the past decades numerous exoplanets – that is, planets belonging to stars other than the Sun – have been discovered. The statistical analysis of their orbits suggests that variants of Bode’s law may roughly hold in other solar systems as well.

1.3 Kepler's Laws of Planetary Motion

In 1543, the year of his death, the Polish astronomer Nicolaus Copernicus published a revolutionary work in which he asserted that the Earth was not the center of the universe. According to his model the Earth rotated about its own axis, and it and the other planets revolved about the Sun. Copernicus calculated the sidereal period of each planet about the Sun; this is the time required for a planet to make one revolution and return to the same angular position relative to a fixed star. He also determined the radii of their orbits about the Sun in terms of the Earth–Sun distance. The mean radius of the Earth's orbit about the Sun is called an *astronomical unit* (AU); it equals 149,597,871 km. Accurate values of these parameters were calculated from observations compiled during an interval of 20 years by the Danish astronomer Tycho Brahe (1546–1601). On his death, the records passed to his assistant, Johannes Kepler (1571–1630). Kepler succeeded in fitting the observations into a heliocentric model for the system of known planets. The three laws in which Kepler summarized his deductions were later to prove vital to Isaac Newton for verifying the law of universal gravitation (Section 3.2.1). It is remarkable that the data base used by Kepler was founded on observations that were unaided by the telescope, which was not invented until early in the seventeenth century.

Kepler took many years to fit the observations of Tycho Brahe into three laws of planetary motion. The first and second laws (Fig. 1.4) were published in 1609 and the third law appeared in 1619. The laws may be formulated as follows:

1. The orbit of each planet is an ellipse with the Sun at one focus.
2. The orbital radius of a planet sweeps out equal areas in equal intervals of time.
3. The ratio of the square of a planet's period (T^2) to the cube of

the semi-major axis of its orbit (a^3) is a constant for all the planets, including the Earth.

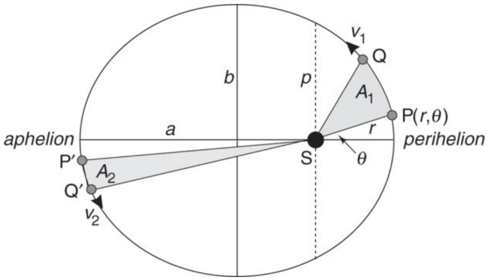


Fig. 1.4 Kepler's first two laws of planetary motion: (1) each planetary orbit is an ellipse with the Sun at one focus, and (2) the radius to a planet sweeps out equal areas in equal intervals of time.

Kepler's three laws were purely empirical, derived from accurate observations. In fact they are expressions of more fundamental physical laws. The elliptical shapes of planetary orbits (Box 1.1) described by the first law are a consequence of the *conservation of energy* of a planet orbiting the Sun under the effect of a central attraction that varies as the *inverse square* of distance. The second law describing the rate of motion of the planet around its orbit follows directly from the *conservation of angular momentum* of the planet. The third law results from the balance between the force of gravitation attracting the planet toward the Sun and the centrifugal force away from the Sun due to its orbital speed. The third law is easily proved for circular orbits (see Section 3.3.2.3).

Kepler's laws were developed for the solar system but are applicable to any closed planetary system. They govern the motions of any natural and artificial satellite about a parent body. Kepler's third law relates the period (T) and the semi-major axis (a) of the orbit of the satellite to the mass (M) of the parent body through the equation

$$GM = \frac{4\pi^2}{T^2}a^3, \tag{1.6}$$

where G is the gravitational constant. This relationship was extremely important for determining the masses of those planets that have natural satellites. It can now be applied to determine the masses of planets using the orbits of artificial satellites.

Special terms are used in describing elliptical orbits. The nearest and furthest points of a planetary orbit around the Sun are called *perihelion* and *aphelion*, respectively. The terms *perigee* and *apogee* refer to the corresponding nearest and furthest points of the orbit of the Moon or a satellite about the Earth.

Box 1.1 Orbital Parameters

The orbit of a planet or comet in the solar system is an ellipse with the Sun at one of its focal points. This condition arises from the conservation of energy in a force-field obeying an inverse square law. The total energy (E) of an orbiting mass is the sum of its kinetic energy (K) and potential energy (U). For an object with mass m and velocity v in orbit at distance r from the Sun (mass S):

$$\frac{1}{2}mv^2 - G\frac{mS}{r} = E = \text{constant.} \tag{1}$$

If the kinetic energy is greater than the potential energy of the gravitational attraction to the Sun ($E > 0$), the object will escape from the solar system. Its path is a *hyperbola*. The same case results if $E = 0$, but the path is a parabola. If $E < 0$, the gravitational attraction binds the object to the Sun; the path is an ellipse with the Sun at one focal point (Fig. B1.1.1). An ellipse is defined as the locus of all points in a plane whose distances s_1 and s_2 from two fixed points F_1 and F_2 in the plane have a constant sum, defined as $2a$:

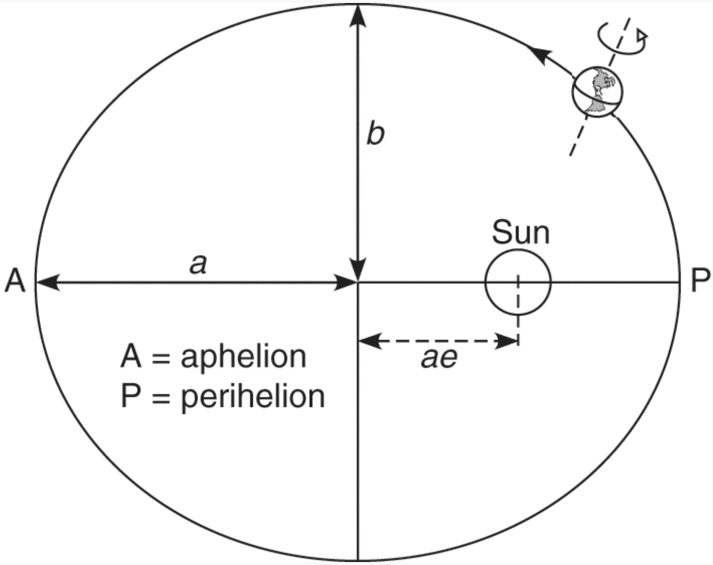


Fig. B1.1.1 The parameters of an elliptical orbit.

$$s_1 + s_2 = 2a. \tag{2}$$

The distance 2a is the length of the major axis of the ellipse; the minor axis perpendicular to it has length 2b, which is related to the major axis by the eccentricity of the ellipse, e :

$$e = \sqrt{1 - \frac{b^2}{a^2}}. \tag{3}$$

The equation of a point on the ellipse with Cartesian coordinates (x, y) defined relative to the center of the figure is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1. \tag{4}$$

The elliptical orbit of the Earth around the Sun defines the ecliptic plane. The angle between the orbital plane and the ecliptic is called the inclination of the orbit, and for most planets except Mercury (inclination 7°) and Pluto (inclination 17°) this is a small angle. A line perpendicular to the ecliptic defines the North and South ecliptic poles. If the fingers of one's right hand are wrapped around the Earth's orbit in the direction of motion, the thumb points to the North ecliptic pole, which is in the constellation Draco ("the dragon"). Viewed from above this pole, all planets move around the Sun in a counterclockwise (prograde) sense.

The rotation axis of the Earth is tilted away from the perpendicular to the ecliptic, forming the angle of obliquity (Fig. B1.1.2), which is currently 23.5°. The equatorial plane is tilted at the same angle to the ecliptic, which it intersects along the line of equinoxes. During the annual motion of the Earth around the Sun, this line twice points to the Sun: on March 20, defining the vernal (spring) equinox, and on September 23, defining the autumnal equinox. On these dates, day and night have equal length everywhere on the Earth. The summer and winter solstices occur on June 21 and December 22, respectively, when the apparent motion of the Sun appears to reach its highest and lowest points in the sky.

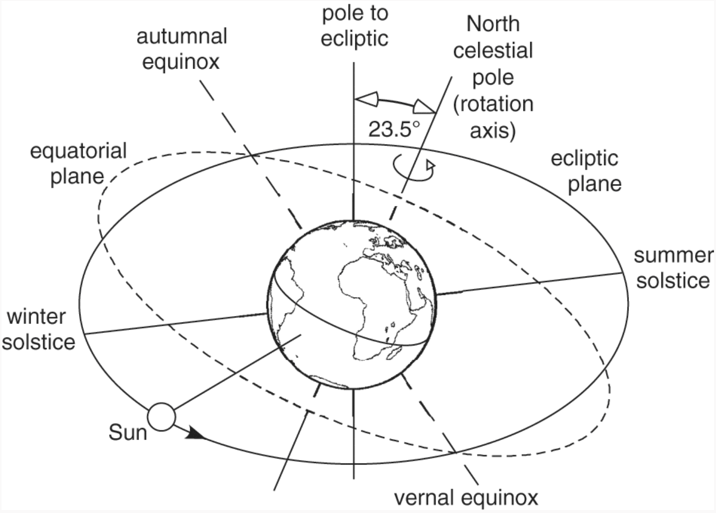

pole to ecliptic
autumnal equinox
North celestial pole (rotation axis)
23.5°
equatorial plane
ecliptic plane
summer solstice
winter solstice
Sun
vernal equinox

Fig. B1.1.2 The relationship between the ecliptic plane, the Earth's equatorial plane, and the line of equinoxes.

1.4 Characteristics of the Planets and Their Orbits

Galileo Galilei (1564–1642) is often regarded as a founder of modern science. He made fundamental discoveries in astronomy and physics, including the formulation of the laws of motion. He was one of the first scientists to use the telescope to acquire more detailed information about the planets. In 1610 Galileo discovered the four largest satellites of Jupiter (Io, Europa, Ganymede, and Callisto), and observed that (like the Moon) the planet Venus exhibited different phases of illumination, from full disk to partial crescent. This was persuasive evidence in favor of the Copernican view of the solar system.

In 1686 Newton applied his theory of universal gravitation to observations of the orbit of Callisto and calculated the mass of Jupiter (J) relative to that of the Earth (E). The value of the gravitational constant G was not yet known; it was first determined by Lord Cavendish in 1798. However, Newton calculated the value of GJ to be 124,400,000 km^3 s^{-2}. This was a very good determination; the modern value for GJ is 126,712,767 km^3 s^{-2}. Observations of the Moon's orbit about the Earth showed that the value GE was 398,600 km^3 s^{-2}. Hence Newton inferred the mass of Jupiter to be more than 300 times that of the Earth.

In 1781 William Herschel discovered Uranus, the first planet to be found by telescope. The orbital motion of Uranus was observed to have inconsistencies, and it was inferred that the anomalies were due to the perturbation of the orbit by a yet undiscovered planet. The predicted new planet, Neptune, was discovered in 1846. Although Neptune was able to account for most of the anomalies of the orbit of Uranus, it was subsequently realized that small residual anomalies remained. In 1914 Percival Lowell predicted the existence of an even more distant planet, the search for which culminated in the detection of Pluto in 1930.

The masses of the planets can be determined by applying Kepler's third law to the observed orbits of natural and artificial satellites and to the tracks of passing spacecraft. Estimation of the sizes and shapes of the planets depends on data from several sources. Early astronomers used occultations of the stars by the planets; an occultation is the eclipse of one celestial body by another, such as when a planet passes between the Earth and a star. The duration of an occultation depends on the diameter of the planet, its distance from the Earth, and its orbital speed.

The dimensions of the planets and the Moon (Table 1.1) have been determined with high precision in modern times using radar-ranging and laser-ranging techniques (Box 1.2). The measurements have been made from instruments located on the Earth and on spacecraft that have orbited or passed close to the target. Both ranging methods measure the two-way travel-time of a pulse of electromagnetic waves, which travels at the speed of light. Radar-ranging operates at wavelengths in the centimeter range while laser-ranging wavelengths are around a micrometer, including the visible range.

Box 1.2 Radar and Laser Ranging

The name *radar* derives from the acronym for **RA**dio **D**etection **A**nd **R**anging, a defensive system developed during World War II for the location of enemy aircraft. A pulsed electromagnetic signal in the microwave frequency range (see Fig. 10.21) is transmitted toward a target, from which a fraction of the incident energy is reflected to a receiver. The laws of optics for visible light apply equally to radar waves, which are subject to reflection, refraction, and diffraction. Visible light has short wavelengths (400–700 nm) and is scattered by the atmosphere, especially by clouds. Radar signals have longer wavelengths

(~1 cm to 30 cm) and can pass through clouds and the atmosphere of a planet with little dispersion. The radar signal is transmitted in a narrow beam of known azimuth, so that the returning echo allows exact location of the direction to the target. The signal travels at the speed of light so the distance, or range, to the target may be determined from the time difference at the source between the transmitted and reflected signals.

The transmitted and reflected radar signals lose energy in transit due to atmospheric absorption, but more importantly, the amplitude of the reflected signal is further affected by the nature of the reflecting surface. Each part of the target's surface illuminated by the radar beam contributes to the reflected signal. If the surface is inclined obliquely to the incoming beam, little energy will reflect back to the source. The reflectivity and roughness of the reflecting surface determine how much of the incident energy is absorbed or scattered. The intensity of the reflected signal can thus be used to characterize the type and orientation of the reflecting surface, e.g., whether it is bare or forested, flat or mountainous.

Laser is the acronym for **L**ight **A**mplification by **S**timulated **E**mission of **R**adiation, and is descriptive of the way in which a monochromatic source of light is produced. Some laser "light" is visible – e.g., with wavelength ~550 nm (red) or ~650 nm (green) – but the range of wavelengths for special applications extends from 200 nm (ultraviolet) to greater than 1000 nm (infrared). A monochromatic laser beam has a single frequency and the light vibrations are in phase. This enables the beam to be kept very narrow so that it can be focused on a small target. For example, a laser beam directed at the Moon in the Lunar Laser Ranging experiment (LRR) is estimated to be 6.5 km wide when it reaches the target on the Moon's surface at a mean distance of 385,000 km from the Earth. The short wavelength of laser light provides very high resolution of the distance to a target. Measurements at a global network of observation stations allows tracking of Earth-orbiting satellites with millimeter-level precision.

The rate of rotation of a planet about its own axis can be determined by observing the motion of features fixed to its surface. For example, radar beams from powerful transmitters on the Earth may be directed at a chosen planet. The two-way travel time gives the distance to the planet. However, the axial rotation of the planet means that one side is rotating away from the Earth and the other side toward the Earth. The Doppler effect (Box 1.3) shifts the frequency of the echo from the receding edge of the planet toward lower frequencies and shifts the echo from the advancing edge toward higher frequencies. The frequency shifts yield the rate of rotation of the planet. In 1965 the 3:2 resonance between the axial rotation rate and orbital rate of Mercury was discovered by this means.

Box 1.3 Doppler Radar

The Doppler effect, first described in 1842 by an Austrian physicist, Christian Doppler, explains how the relative motion between source and detector influences the observed frequency of light and sound waves. For example, suppose a stationary radar source emits a signal consisting of n_0 pulses per second. The frequency of pulses reflected from a stationary target at distance d is also n_0, and the two-way travel time of each pulse is equal to $2(d/c)$, where c is the speed of light. If the target is moving toward the radar source, its velocity v shortens the distance between the radar source and the target by $(vt/2)$, where t is the new two-way travel time:

$$t = 2\left(\frac{d - (vt/2)}{c}\right) = t_0 - \frac{v}{c}t,$$

(1)

$$t = \frac{t_0}{(1 + v/c)}.$$

(2)

The travel time of each reflected pulse is shortened, so the number of reflected pulses (*n*) received per second is correspondingly higher than the number emitted:

$$n = n_0\left(1 + \frac{v}{c}\right).$$

(3)

The opposite situation arises if the target is moving away from the source: the frequency of the reflected signal is lower than that emitted. Similar principles apply if the radar source is mounted on a moving platform, such as an aircraft or satellite. The Doppler change in signal frequency in each case allows remote measurement of the relative velocity between an object and a radar transmitter.

In another important application, the Doppler effect provides evidence that the universe is expanding. The observed frequency of light from a star (i.e., its color) depends on the velocity of its motion relative to an observer on the Earth. The color of the star shifts toward the red end of the spectrum (lower frequency) if the star is receding from the Earth and toward the blue end (higher frequency) if it is approaching the

Earth. The color of light from many distant galaxies has a “red shift,” implying that these galaxies are receding from the Earth.

Radar-ranging in combination with Doppler analysis has been used to measure the distances and rotation rates of the planets as far as Saturn, as well as many asteroids and the Moon. Where this is not possible (e.g., the surface of Uranus is featureless) other techniques must be employed. In the case of Uranus the rotational period of 17.2 hours was determined from periodic radio emissions produced by electrical charges trapped in its magnetic field; they were detected by the Voyager 2 spacecraft when it flew by the planet in 1986.

All the planets revolve around the Sun in the same sense, which is counterclockwise when viewed from above the plane of the Earth’s orbit (called the *ecliptic* plane). Except for Pluto, the orbital plane of each planet is inclined to the ecliptic at a small angle (Table 1.2). Most of the planets rotate about their rotation axis in the same sense as their orbital motion about the Sun, which is termed *prograde*. Venus rotates in the opposite, *retrograde* sense. The angle between a rotation axis and the ecliptic plane is called the *obliquity* of the axis. The rotation axes of Uranus and Pluto lie close to their orbital planes. Each is tilted away from the pole to the orbital plane at an angle greater than 90°, so that, strictly speaking, their rotations are also retrograde.

1.5 The Inner (Terrestrial) Planets and the Moon

The inner planets of the solar system are those that orbit closest to the Sun. Known as the terrestrial planets because of their Earth-like properties, they typically have a dense metallic core within a rocky silicate body. Their hard solid surfaces are characterized by topographic features such as valleys, mountains and craters. They have atmospheres that were acquired during their evolution rather than from their original formation, as is the case for the outer planets. None of the inner planets has a system of rings. Mercury and Venus have no natural satellites, and Mars has two small, irregularly shaped moons less than 10 km in diameter. The Earth's natural satellite, the Moon, is the largest in the solar system in relation to the planet it orbits.

1.5.1 Mercury

Mercury is the closest planet to the Sun. This proximity and its small size make it difficult to study Mercury telescopically. Its orbit has a large eccentricity (0.2056). At perihelion the planet is about 46.0 million km (0.3075 AU) from the Sun, but at aphelion the distance is 69.8 million km (0.4667 AU). Until 1965 the rotational period was thought to be the same as the period of revolution (88 days), so that it would keep the same face to the Sun, in the same way that the Moon does to the Earth. However, in 1965 Doppler radar measurements showed that this is not the case. In 1974 and 1975 images from the close passage of Mariner 10, the first spacecraft to have visited the planet, gave a period of rotation of 58.8 days, and Doppler tracking gave a radius of 2439 km; both values are close to the presently accepted ones. In 2004 NASA launched the unmanned MESSENGER spacecraft on a mission to measure Mercury's atmosphere, gravity, geology, and magnetic field. MESSENGER's trajectory involved fly-bys of Venus, the Earth, and Mercury (in 2008), before the spacecraft entered orbit around Mercury in 2011. It subsequently orbited the planet continuously until 2015, when its fuel expired and it crashed into the planet, terminating the mission. The mission yielded improved measurements of the physical properties of the planet (Table 1.1) and its magnetic field (Table 11.1).

The spin and orbital motions of Mercury are both prograde and are coupled in the ratio 3:2. The spin period is 58.65 Earth days, almost exactly 2/3 of its orbital period of 87.969 Earth days. For an observer on the planet this has the unusual consequence that a Mercury day lasts longer than a Mercury year! During one orbital revolution about the Sun (one Mercury year) an observer on the surface rotates about the spin axis 1.5 times and thus advances by an extra half-turn. If the Mercury year started at sunrise, it would end at sunset, so the observer on Mercury would spend the entire 88 Earth days exposed to solar heating, which causes the surface temperature to exceed 700 K. During the following Mercury year, the axial rotation advances by a further half-turn, during which the observer is on the night side of the planet for 88 days, and the temperature sinks below 100 K. After two solar orbits and three axial rotations, the observer is back at the starting point. The range of temperatures on the surface of Mercury is the most extreme in the solar system.

Although the mass of Mercury is only about 5.5% that of the Earth, its mean radius of 2439.7 km is much smaller, so that its mean density of 5427 kg m^{-3} is comparable to Earth's (5514 kg m^{-3}) and is the second highest in the solar system. Mercury's interior is dominated by a proportionately large iron core with a radius of 2020 km. As on the Earth, the outer core is liquid and there may be a solid inner core. The partly molten core is enclosed in a thin solid outer shell, equivalent to the Earth's mantle and crust. Mercury has a weak, nearly axial, dipole magnetic field; its center is offset by several hundred kilometers along the rotation axis northward from the planet's equator. Its proximity to the Sun subjects Mercury to intense illumination and a strong solar wind, against which the weak magnetic field provides ineffective shielding. Consequently, much of the planet's atmosphere is dispersed downwind.

1.5.2 Venus

Venus is the brightest object in the sky after the Sun and Moon. Its orbit brings it closer to the Earth than any other planet, which made it an early object of study by telescope. Its occultation with the Sun was observed telescopically as early as 1639. Estimates of its radius based on occultations indicated about 6120 km. Galileo observed that the apparent size of Venus changed with its position in orbit and, like the Moon, the appearance of Venus showed different phases from crescent-shaped to full. This was important evidence in favor of the Copernican heliocentric model of the solar system, which had not yet replaced the Aristotelian geocentric model.

Venus has the most nearly circular orbit of any planet, with an eccentricity of only 0.00677 and mean radius of 0.7233 AU (Table 1.2). Its orbital period is 224.7 Earth days, and the period of rotation about its own axis is 243.0 Earth days; thus a Venusian day lasts longer than a Venusian year. Its spin axis is tilted at 177.4° to the pole to the ecliptic, thus making its spin *retrograde*. The combination of these motions results in the length of a day on Venus (the time between successive sunrises on the planet) being equal to about 117 Earth days.

Venus is very similar in size and probable composition to the Earth. During a near-crescent phase the planet is ringed by a faint glow indicating the presence of an atmosphere. This has been confirmed by several spacecraft that have visited the planet since the first visit by Mariner 2 in 1962. From 2006 until 2014 the Venus Express spacecraft orbited Venus, mainly exploring the atmosphere and observing atmospheric lightning. The atmosphere consists of 96.5% carbon dioxide and 3.5% nitrogen, and is very dense. This results in a surface atmospheric pressure that is 92 times that of the Earth. Thick cloud cover results in a strong greenhouse effect that produces stable temperatures up to 740 K, slightly higher than the maximum day-time values on Mercury, making Venus the hottest of the planets. The thick clouds obscure any view of the surface, which has however been surveyed with radar. In 1967 the Soviet spacecraft Venera 4 explored the composition of the Venusian atmosphere and landed a probe on the surface of the planet. The Magellan spacecraft, which was placed in a nearly polar orbit around the planet in 1990, carried a radar-imaging system with an optimum resolution of 100 meters, and a radar altimeter system to measure the topography and some properties of the planet's surface.

Venus is unique among the planets in rotating in a *retrograde* sense about an axis that is almost normal to the ecliptic. (Note: Uranus can also be considered to have retrograde rotation, but its axis lies close to the ecliptic.) Like Mercury, Venus has a high Earth-like density (5243 kg m^{-3}). On the basis of its density together with gravity estimates from orbiting and fly-by spacecraft, it is thought that the interior of Venus may be similar to that of the Earth, with a rocky mantle surrounding an iron core about 3000 km in radius, that is possibly at least partly molten. However, in contrast to the Earth, Venus does not have an intrinsic magnetic field.

1.5.3 The Earth

The Earth moves around the Sun in a slightly elliptical orbit. The parameters of the orbital motion are important, because they define astronomical units of distance and time. The Earth’s rotation about its own axis from one solar zenith to the next one defines the solar day (see Section 8.1.2). The length of time taken for it to complete one orbital revolution about the Sun defines the solar year, which is equal to 365.242 solar days. The eccentricity of the orbit is presently 0.01671, but it varies between a minimum of 0.001 and a maximum of 0.060 with a period of about 100,000 years due to the influence of the other planets. The mean radius of the orbit (149,597,871 km) is called an *astronomical unit* (AU). Distances within the solar system are usually expressed as multiples of this unit. The distances to extra-galactic celestial bodies are expressed as multiples of a light-year (the distance traveled by light in a vacuum in one year). The Sun’s light takes about 8 min 20 s to reach the Earth. Due to the difficulty of determining the gravitational constant, the mass of the Earth (E) is not known with high precision, but is estimated to be 5.9723×10^{24} kg. In contrast, the product GE is known accurately; it is equal to $3.986004418 \times 10^{14}\ \mathrm{m^3\ s^{-2}}$. The rotation axis of the Earth is presently inclined at 23.439° to the pole of the ecliptic. However, the effects of other planets also cause the angle of *obliquity* to vary between a minimum of 21.9° and a maximum of 24.3°, with a period of about 41,000 years.

1.5.4 The Moon

The Moon is the Earth's only natural satellite. Its equatorial radius of 1738 km makes it much larger relative to its parent body than the natural satellites of the other planets, except for Pluto's moon, Charon. There have been many attempts to explain the Moon's origin, which is still poorly understood. The best known is the "Giant Impact" hypothesis. This maintains that about 4.5 billion years (Gyr) ago, shortly after the solar system formed, the Earth collided with a postulated Mars-sized planet, which has been given the name *Theia*. Parts of Theia accreted into the early Earth and the debris from the collision subsequently coalesced to form the Moon. The Giant Impact hypothesis satisfies many properties of the Earth–Moon system, and reflects the currently prevailing opinion in planetary science. Nevertheless, other lunar origin models have been proposed – for example, that Earth and Moon accreted together, or that the Moon was captured during a close encounter with the early Earth.

The Moon's distance from the Earth was first estimated with the method of parallax. Instead of observing the Moon from different positions of the Earth's orbit, as in Fig. 1.2, the Moon's position relative to a fixed star was observed at times 12 hours apart, close to moonrise and moonset, when the Earth had rotated through half a revolution. The baseline for the measurement is then the Earth's diameter. The distance of the Moon from the Earth was found to be about 60 times the Earth's radius.

Modern investigations of the Moon have been carried out from the Earth and from orbiting spacecraft. Laser-ranging (Section 3.5) delivers very accurate measurements of the Earth–Moon distance, which has been monitored in the LLR experiment since 1969. Retroreflector panels (Box 1.4) were installed at five locations on the Moon by astronauts of the American Apollo 11, 14, and 15 missions, and by remote landers of the Soviet Union's Luna missions. A beam of rapid, short laser pulses transmitted from the Earth is reflected by the reflectors exactly back along its path. The two-way travel-time of a pulse is measured accurately; it travels at the speed of light and so the distance of a reflector from the Earth can be determined. The data from a global network of geodetic stations is consolidated to give the Earth–Moon distance with millimeter-level accuracy. Consequently the parameters of the Moon's orbit around the Earth are very well known. The Moon's orbit is slightly elliptical, with eccentricity 0.0549 and mean radius 384,100 km.

Box 1.4 Principle of the Retroreflector

A retroreflector is a device that reflects a light ray along a path that is exactly opposite to its incident path. Two types are in common use: cube-corner and spherical retroreflectors. The cube-corner type makes use of the law of optical reflection, which states that the angles between the normal to the reflecting surface and the incident and reflected rays are equal. Consider an incident light ray with components *h* parallel and *v* perpendicular to the reflecting surface as in Fig. B1.4.1a. The horizontal component is the same for both rays but the component *v* normal to the reflector changes sign.

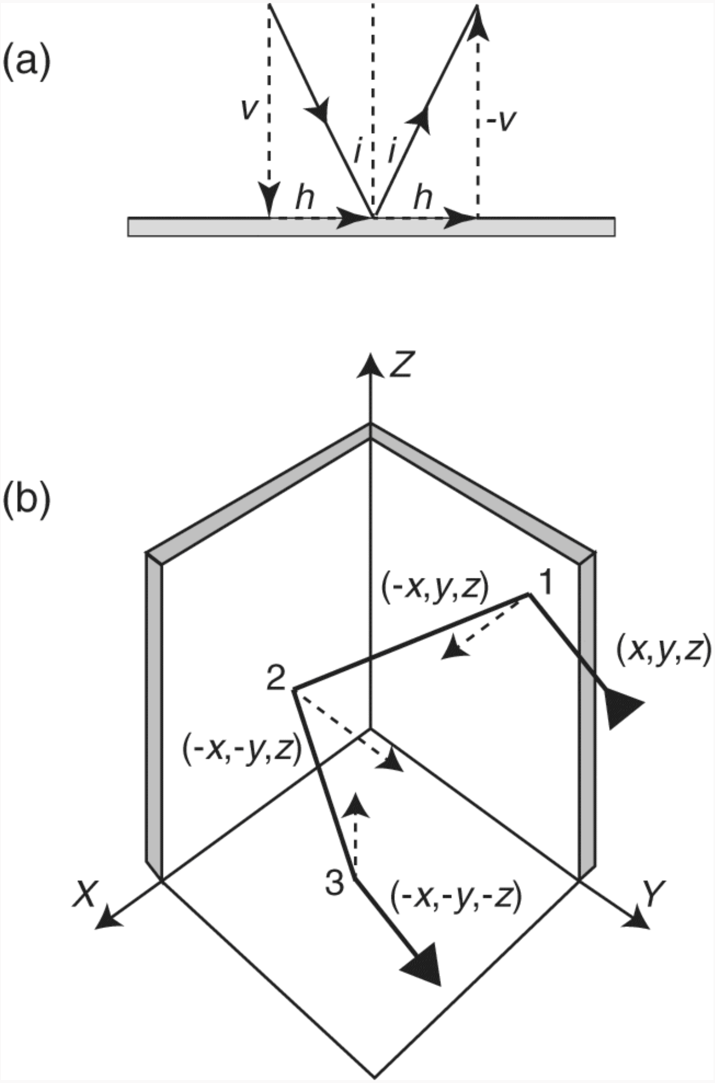


Fig. B1.4.1 The path of a light ray in a corner retroreflector.

A cube-corner reflector consists of three mutually orthogonal reflecting planes, such as the corner of a cube with silvered external surfaces. Three mirrors at right angles to each other form an equivalent system. A light ray of light incident on the first mirror is reflected in turn from each of the other mirrors. Let (*x, y, z*) be the components of the incident ray (Fig. B1.4.1b) with respect to the orthogonal axes (*X, Y, Z*). After reflection at mirror 1 the *x*-component changes sign, and the ray has components (–*x*, *y*, *z*); after reflection 2 the components are (–*x*, –*y*, *z*); and after reflection 3 the light ray has components (–*x*, –*y*, –*z*) and has completely reversed its direction. It leaves the retroreflector in exactly the opposite direction to the incident ray.

A spherical retroreflector (Fig. B1.4.2) makes use of the law of refraction as well as the law of reflection. An incident ray is refracted (bent) at the front surface of a transparent sphere, which acts like a lens that has the silvered rear surface of the sphere as its focal surface. The refractive index of the glass sphere is chosen to meet this condition. The incident angle *r* on the reflecting surface is the same as the refracted angle on entering the sphere at its front surface. The reflected ray passes out of the sphere with a further refraction at its outer surface, so that the return path of the ray is exactly opposite to the direction of its arrival. This type of retroreflector is familiar to the "cat's eyes" that serve to mark the center or margins of a road at night or in fog. They also have numerous other uses in reflective materials.

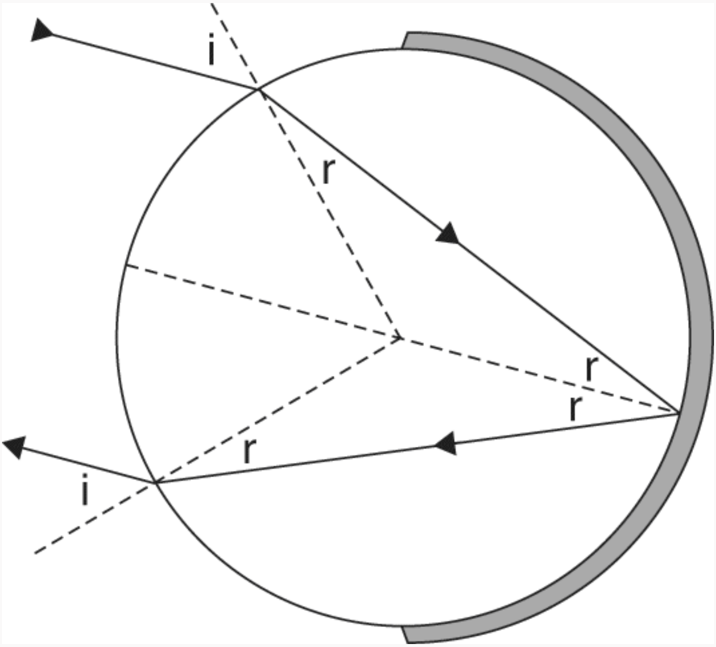


Fig. B1.4.2 The path of a light ray in a spherical retroreflector.

The Moon rotates about its axis in the same sense as its orbital revolution about the Earth. Tidal friction resulting from the Earth's attraction has slowed down the Moon's rotation, so that it now has the same mean period as its revolution, 27.32 days. As a result, the Moon always presents the same face to the Earth. In fact, slightly more than half (about 59%) of the lunar surface can be viewed from the Earth. Several factors contribute to this. First, the plane of the Moon's orbit around the Earth is inclined at 5°9′ to the ecliptic while the Moon's equator is inclined at 1°32′ to the ecliptic. The inclination of the Moon's equator varies by up to 6°41′ to the plane of its orbit. This is called the *libration of latitude*. It allows Earth-based astronomers to see 6°41′ beyond each of the Moon's poles. Second, the Moon moves with variable velocity around its elliptical orbit, while its own rotation is constant. Near perigee the Moon's orbital velocity is fastest (in accordance with Kepler's second law) and the rate of revolution exceeds slightly the constant rate of lunar rotation. Similarly, near apogee the Moon's orbital velocity is slowest and the rate of revolution is slightly less than the rate of rotation. The rotational differences are called the Moon's *libration of longitude*. Their effect is to expose zones of longitude beyond the average edges of the Moon. Finally, the Earth's diameter is quite large compared to the Moon's distance from Earth. During Earth's rotation the Moon is viewed from different angles that allow about one additional degree of longitude to be seen at the Moon's edge.

The gravitational attraction of the Earth and the Moon on each other slows the rates of rotation of both bodies. The loss of angular momentum is compensated by an equivalent gain in the orbital angular momentum of the Moon, achieved by an increase of the radius of the Moon's orbit. The LLR experiment has determined that the Moon is moving away from the Earth at a mean rate of 37.9 mm yr^{-1}. The high accuracy of LLR observations allows measurement of the tides induced in the Moon by the Earth's gravitation, and these have been interpreted in terms of the Moon's internal structure. They suggest that the Moon has a small, dense core.

Since the 1960s several spacecraft have orbited the Moon to investigate its magnetic and gravitational fields. The Moon was found to have no global magnetic field, but the presence of quite strongly magnetized rocks suggest it may have had one early in its history (Section 11.3.1.2). The orbiting spacecraft measured the oblateness of the Moon's shape and discovered subsurface mass concentrations ("mascons") of unknown origin. In 2012 the Moon's gravitational field was mapped in great detail by the **G**ravity **R**ecovery **A**nd **I**nterior **L**aboratory (GRAIL) mission. Based on the technology of the earlier GRACE investigation of the Earth's gravitational field (Section 3.5.4), the GRAIL mission employed two identical spacecraft, which orbited the Moon in formation in the same low-altitude, nearly polar orbit. A microwave monitoring system on board each spacecraft measured changes in their separation with a precision of one micrometer. When the leading spacecraft crossed a lunar mass anomaly it speeded up, or slowed down, changing the separation. This change was repeated when the following spacecraft crossed the mass anomaly. The GRAIL mission was active for nine months, during which it mapped the Moon's gravitational field in unparalleled detail. The mathematical description of the lunar potential by spherical harmonics (Box 3.4) encompasses hundreds of terms, so that the potential of the lunar topography can be matched with high precision. In particular, mascons were located and described more exactly. GRAIL showed that the gravity-high of a mascon is surrounded by a low-gravity ring, as would result from the impact of an asteroid, which could have occurred earlier in the Moon's history.

The continuous, detailed monitoring in the LLR experiment and the high-resolution gravitational field measured by the GRAIL mission indicate that the Moon has a differentiated internal structure, with a thin crust overlying a silicate mantle, and a small central core. More detailed analysis of the internal structure is obtained from seismological evidence gathered by seismometers deployed at four locations on the Moon between 1969 and 1972 by Apollo astronauts, which continued to operate until 1977. The recorded seismicity consisted of “moonquakes” with body-wave magnitudes (Section 7.2.6.2) mostly smaller than 3, originating in a number of deep sources primarily at depths of 700–1200 km. On the Earth, most earthquakes result from tectonic processes, but there are no tectonic plate motions on the Moon. Instead, the monthly interval of moonquake occurrence suggests that they are caused by internal stresses related to the tides raised in the body of the Moon by the Earth and Sun. Although the data basis from only four locations is limited, modern seismic array processing of lunar seismicity revealed consistent reflecting discontinuities at depth (Fig. 1.5). They correspond to an internal structure in which a small, solid, iron core with radius 240 km is surrounded by a fluid core extending to 330 km radius. The lunar core is proportionately very much smaller than the Earth’s in relation to the size of the body. Above the core is a thick silicate mantle and a thin crust that is variously estimated to be 40–60 km thick. The lowest part of the mantle – between 330 km and 480 km radii – may be partially melted. Because of the small size of the core, the density structure associated with the seismic layering yields a moment of inertia for the Moon that is only slightly less than if the Moon had a uniform density.

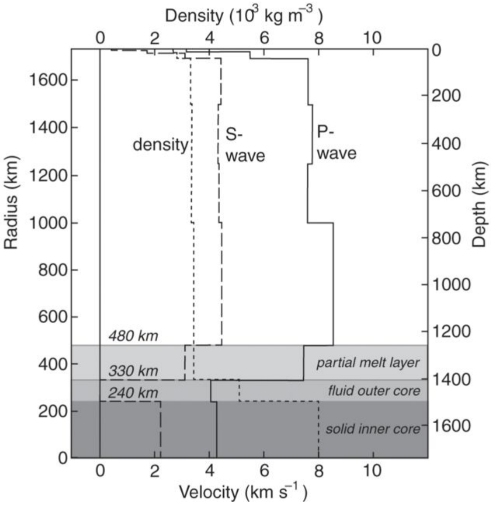


Fig. 1.5 The internal structure of the Moon: concentric shells with different compositions and states are shown by the depth profiles of seismic P-wave and S-wave velocities and density.

Source: R.C. Weber et al., Seismic detection of the lunar core, *Science* 331, 309–312, 2011.

1.5.5 Mars

Mars, popularly called the red planet because of its hue when viewed from Earth, has been known since prehistoric times and was also an object of early telescopic study. In 1666 Gian Domenico Cassini determined the rotational period at just over 24 hours; radio-tracking from two Viking spacecraft that landed on Mars in 1976, more than three centuries later, gave a period of 24.624 hours. Subsequently, numerous spacecraft have orbited Mars or landed on its surface. Although much is now known about the planet, the structure and dynamics of its interior are still poorly understood.

The orbit of Mars is quite elliptical (eccentricity 0.0935). The large difference between perihelion and aphelion causes large temperature variations on the planet. The average surface temperature is about 218 K, but temperatures range from 140 K at the poles in winter to 300 K on the day side in summer. Mars has two natural satellites, Phobos and Deimos. Observations of their orbits gave early estimates of the mass of the planet. Its size was established quite early telescopically from occultations. Its shape is known very precisely from spacecraft observations. The polar flattening is about double that of the Earth. The rotation rates of Earth and Mars are almost the same, but the lower mean density of Mars results in smaller gravitational forces, so at any radial distance the relative importance of the centrifugal acceleration is larger on Mars than on Earth.

In 2004 the Mars Expedition Rover vehicles *Spirit* and *Opportunity* landed on Mars and transmitted photographs and geological information to Earth. Three spacecraft (*Mars Global Surveyor, Mars Odyssey*, and *Mars Express*) were placed in orbit to carry out surveys of the planet. These and earlier orbiting spacecraft and Martian landers have revealed details of the planet that cannot be determined with a distant telescope (including the Earth-orbiting Hubble telescope). Much of the Martian surface is very old and cratered, but there are also much younger rift valleys, ridges, hills, and plains. The topography is varied and dramatic, with mountains that rise to 24 km, a 4000 km long canyon system, and impact craters up to 2000 km across and 6 km deep.

The internal structure of Mars has been inferred from the results of these missions. Mars has a relatively low mean density (3933 kg m^{-3}) compared to the other terrestrial planets. Its mass is only about one-tenth that of Earth (Table 1.1), so the pressures in the planet are lower and the interior is less densely compressed. Mars has an internal structure similar to that of the Earth. A thin crust, 35 km thick in the northern hemisphere and 80 km thick in the southern hemisphere, surrounds a rocky mantle whose rigidity decreases with depth as the internal temperature increases. The planet has a dense core 1500–1800 km in radius, thought to be composed of iron with a relatively large fraction of sulfur. Minute perturbations of the orbit of *Mars Global Surveyor*, caused by deformations of Mars due to solar tides, have provided more detailed information about the internal structure. They indicate that, like the Earth, Mars probably has a solid inner core and a liquid outer core that is, however, too small to generate a global magnetic field.

The global mapping mission of *Mars Global Surveyor* ended in 2006, having delivered important results that raised new questions about the planet's structure and history. Since August 2012 the largely autonomous Mars rover *Curiosity* has investigated the surface and atmosphere of Mars using a wide range of spectrographs, cameras, and meteorological instruments. In November 2018, NASA's InSight mission placed various geophysical instruments on the surface of Mars to evaluate the planet's seismicity, and to measure the surface heat flow. The lander will also make precise observations of the nutation of the planet's rotation axis (see Section 3.3.5), which will deliver additional information about the Martian core and mantle.

1.5.6 The Asteroids

The asteroids occur in many sizes, ranging from several hundred kilometers in diameter, down to small bodies that are too small to discern from Earth. There are 26 asteroids larger than 200 km in diameter, but there are probably more than one million with diameters around 1 km. Some asteroids have been photographed by spacecraft in fly-by missions: In 1997 the NEAR-Shoemaker spacecraft orbited and landed on the asteroid Eros. Hubble Space Telescope imagery has revealed details of Ceres (diameter 950 km), Pallas (diameter 830 km), and Vesta (diameter 525 km), which suggest that it may be more appropriate to call these three bodies protoplanets (i.e., still in the process of accretion from planetesimals) rather than asteroids. All three are differentiated and have a layered internal structure like a planet, although the compositions of the internal layers are different. Ceres has an oblate spheroidal shape and a silicate core, and is the most massive asteroid; it has recently been reclassified as a “dwarf planet.” Vesta’s shape is more irregular and it has an iron core.

Asteroids are classified by type, reflecting their composition (stony carbonaceous or metallic nickel–iron), and by the location of their orbits. *Main belt* asteroids have near-circular orbits with radii 2–4 AU between Mars and Jupiter. The *Centaur* asteroids have strongly elliptical orbits that take them into the outer solar system. The *Aten* and *Apollo* asteroids follow elliptical Earth-crossing orbits. The collision of one of these asteroids with the Earth would have a cataclysmic outcome. A 1 km diameter asteroid would create a 10 km diameter crater and release as much energy as the simultaneous detonation of most or all of the nuclear weapons in the world’s arsenals. In 1980 Luis and Walter Alvarez and their colleagues showed on the basis of an anomalous concentration of extra-terrestrial iridium at the Cretaceous–Tertiary boundary at Gubbio, Italy, and other locations, that a 10 km diameter asteroid had probably collided with

Earth, causing directly or indirectly the mass extinctions of many species, including the demise of the dinosaurs. There are 240 known Apollo bodies; however, there may be as many as 2000 that are 1 km in diameter and many thousands more measuring tens or hundreds of meters.

Scientific opinion is divided on what the asteroid belt represents. One idea is that it may represent fragments of an earlier planet that was broken up in some disaster. Alternatively, it may consist of material that was never able to consolidate into a planet, perhaps due to the powerful gravitational influence of Jupiter.

1.6 The Outer (Giant) Planets

The giant planets are largely gaseous, consisting mostly of hydrogen and helium, with traces of methane, water and solid matter. Their compositions are inferred indirectly from spectroscopic evidence, because space probes have not penetrated their atmospheres to any great depth. In contrast to the rocky terrestrial planets and the Moon, the radius of a giant planet does not correspond to a solid surface, but is taken to be the level that corresponds to a pressure of 1 bar, which is approximately Earth's atmospheric pressure at sea level.

Each of the giant planets is encircled by a set of concentric rings, made up of numerous particles. The rings around Saturn, discovered by Galileo in 1610, are the most spectacular. For more than three centuries they appeared to be a feature unique to Saturn, but in 1977 discrete rings were also detected around Uranus. In 1979 the Voyager 1 spacecraft detected faint rings around Jupiter, and in 1989 the Voyager 2 spacecraft confirmed that Neptune also has a ring system.

1.6.1 Jupiter

Jupiter has been studied from ground-based observatories for centuries, and more recently with the Hubble Space Telescope, but our detailed knowledge of the planet comes primarily from unmanned space probes that sent photographs and scientific data back to Earth. Between 1972 and 1977 the planet was visited by the Pioneer 10 and 11, Voyager 1 and 2, and Ulysses spacecraft. The spacecraft Galileo orbited Jupiter for eight years, from 1995 to 2003, and sent an instrumental probe into the atmosphere. It penetrated to a depth of 140 km before being crushed by the atmospheric pressure. The Juno spacecraft was inserted into a highly elongate polar orbit around Jupiter in 2016 with the goal of investigating Jupiter's atmosphere and mapping its gravity and magnetic fields. Jupiter's atmosphere is characterized by visible belts of eastward and westward flowing jet streams, which exhibit asymmetry between the north and south hemispheres and extend to depths of 2000–3000 km into the giant planet. Jupiter's gravity field reflects this asymmetry because the deeper the extent of the jets the greater the mass anomaly they cause. Analysis of the gravitational field found that the differential rotation decreases with depth so that the deep interior appears to rotate as a solid. There are conspicuous atmospheric features in the polar regions, where a central cyclone is surrounded by smaller cyclones, which do not merge into each other but remain distinct over time.

Jupiter is by far the largest of all the planets. Its mass (19×10^{26} kg) is 318 times that of the Earth (Table 1.1) and 2.5 times the mass of all the other planets added together (7.7×10^{26} kg). Despite its enormous size the planet has a very low density of only 1326 kg m^{-3}, from which can be inferred that its composition is dominated by hydrogen and helium. Jupiter has at least 79 moons. The four largest – Io, Europa, Ganymede, and Callisto – were discovered in 1610 by Galileo. The orbital motions of Io, Europa, and Ganymede are synchronous, with periods locked in the ratio 1:2:4. In a few hundred million years, Callisto will also become synchronous with a period eight times that of Io. Ganymede is the largest satellite in the solar system; its radius of 2631 km makes it slightly larger than the planet Mercury. It has a differentiated interior with an iron-rich core, mantle, and icy outer layer. Some of the outermost satellites are less than 30 km in radius, revolve in retrograde orbits, and may be captured asteroids. Jupiter has a system of rings, which are like those of Saturn but are fainter and smaller, and were first detected during analysis of data from Voyager 1. Subsequently, they were investigated in detail during the Galileo mission. In contrast to Saturn, which has rings that are visible through terrestrial telescopes and are composed of ice particles, the rings of Jupiter consist of dust formed by the impact of meteorites on Jupiter's moons.

Jupiter has a dense core, but its physical state is uncertain. Gravitational measurements and theories of planetary formation suggest that the core may be solid. This is surrounded by concentric layers of hydrogen, first in a liquid-metallic state (which means that its atoms, although not bonded to each other, are so tightly packed that the electrons can move easily from atom to atom), then non-metallic liquid, and finally a gaseous outer layer. The planet's atmosphere consists of approximately 86% hydrogen and 14% helium, with traces of methane, water, and ammonia. The liquid-metallic hydrogen layer is a good conductor of electrical currents. These are the source of a powerful magnetic field that is many times stronger than the Earth's and is enormous in extent. It stretches for several million kilometers toward the Sun and for several hundred million kilometers away from it. The magnetic field traps charged particles from the Sun, forming a zone of intense radiation outside Jupiter's atmosphere that would be fatal to a human being exposed to it. The motions of the electric charges cause radio emissions. These are modulated by the rotation of the planet and are used to estimate the period of rotation, which is about 9.9 hours.

Jupiter’s moon Europa is the focus of great interest because of the possible existence of water below its icy crust, which is smooth and reflects sunlight brightly. The Voyager spacecraft took high-resolution images of the moon’s surface, and gravity and magnetic data were acquired during close passages of the Galileo spacecraft. Europa has a radius of 1565 km, so is only slightly smaller than Earth’s Moon, and is inferred to have an iron–nickel core within a rocky mantle, and an outer shell of water below a thick surface ice layer.

1.6.2 Saturn

Saturn is the second largest planet in the solar system, with an equatorial radius of 60,268 km. Its mean density of 687 kg m^{-3} is the lowest in the solar system and less than that of water. Thin concentric rings in its equatorial plane give the planet a striking appearance. The obliquity of its rotation axis to the ecliptic is 26.7°, similar to that of the Earth (Table 1.1). Consequently, as Saturn moves along its orbit the rings appear at different angles to an observer on the Earth. Galileo studied the planet by telescope in 1610 but the early instrument could not resolve details and he was unable to interpret his observations as a ring system. The rings were explained by Christiaan Huygens in 1655 using a more powerful telescope. In 1675, Domenico Cassini observed that Saturn's rings consisted of numerous small rings with gaps between them. Not until after 1977 was it established that systems of rings also exist about the other giant planets: Jupiter, Uranus, and Neptune. The rings are composed of particles of ice, rock, and debris, ranging in size from dust particles up to a few cubic meters, which are in orbit around the planet. The origin of the rings is unknown; one theory is that they are the remains of an earlier moon that disintegrated, either due to an extra-planetary impact or as a result of being torn apart by bodily tides caused by Saturn's gravity.

In addition to its ring system Saturn has more than 30 moons, the largest of which, Titan, has a radius of 2575 km and is the only moon in the solar system with a dense atmosphere. Observations of the orbit of Titan allowed the first estimate of the mass of Saturn to be made in 1831. Saturn was visited by the Pioneer 11 spacecraft in 1979 and later by Voyager 1 and Voyager 2. In 2004 the spacecraft Cassini entered orbit around Saturn, and orbited the planet for the following 13 years. During the mission Cassini launched an instrumental probe, Huygens, that landed on Titan in January 2005. Data from the probe were obtained during the descent by parachute through Titan's atmosphere and after landing, and relayed to Earth by the orbiting

Cassini spacecraft.

Saturn's period of rotation has been deduced from modulated radio emissions associated with its magnetic field. The equatorial zone has a period of 10 h 14 min, while higher latitudes have a period of about 10 h 39 min. The shape of the planet is known from occultations of radio signals from the Voyager spacecraft. The rapid rotation and fluid condition result in Saturn having the greatest degree of polar flattening of any planet, amounting to almost 10%. Its mean density of 687 kg m^{-3} (less than that of water) is the lowest of all the planets, implying that Saturn, like Jupiter, is made up mainly of hydrogen and helium and contains few heavy elements. The planet probably also has a similar layered structure, with a rocky core overlain successively by layers of liquid-metallic hydrogen and molecular hydrogen. However, the gravitational field of Jupiter compresses hydrogen to a metallic state, which has a high density. This gives Jupiter a higher mean density than Saturn. Saturn has a planetary magnetic field that is weaker than Jupiter's but probably originates in the same way.

1.6.3 Uranus

Uranus is so remote from the Earth that Earth-bound telescopic observation reveals no surface features. Until the fly-past of Voyager 2 in 1986 the planet's properties had to be surmised indirectly and were inaccurately known. Voyager 2 provided detailed information about the size, mass and surface of the planet and its satellites, and of the structure of the planet's ring system. The planet's radius is 25,559 km and its mean density is 1270 kg m^{-3}. The rotational period, 17.24 h, was inferred from periodic radio emissions detected by Voyager and believed to arise from charged particles trapped in the magnetic field and thus rotating with the planet. The rotation results in a polar flattening of 2.3%. Prior to Voyager, there were 5 known moons. Voyager discovered a further 10 small moons, and subsequently 12 more moons have been discovered at a greater distance from the planet, bringing the total of Uranus' known moons to 27. The composition and internal structure of Uranus are probably different from those of Jupiter and Saturn. The higher mean density of Uranus suggests that it contains proportionately less hydrogen and more rock and ice. The rotation period is too long for a layered structure with melted ices of methane, ammonia, and water around a molten rocky core. It agrees better with a model in which heavier materials are less concentrated in a central core, and the rock, ices, and gases are more uniformly distributed.

Several paradoxes remain associated with Uranus. The axis of rotation is tilted at an angle of 98° to the pole to the planet's orbit, and thus lies close to the ecliptic plane. The reason for the extreme tilt, compared to the other planets, is unknown. The planet has a prograde rotation about this axis. However, if the other end of the rotation axis, inclined at an angle of 82°, is taken as reference, the planet's spin can be regarded as retrograde. Both interpretations are equivalent. The anomalous axial orientation means that during the 84 years of an orbit round the Sun the polar regions as well as the equator experience extreme solar insolation. The magnetic field of Uranus is also anomalous: it is inclined at a large angle to the rotation axis and its center is displaced axially from the center of the planet.

1.6.4 Neptune

Neptune is the outermost of the gaseous giant planets. It can only be seen from Earth with a good telescope. By the early nineteenth century, the motion of Uranus had become well enough charted that inconsistencies were evident. French and English astronomers independently predicted the existence of an eighth planet, and the predictions led to the discovery of Neptune in 1846. The planet had been noticed by Galileo in 1612, but due to its slow motion he mistook it for a fixed star. The period of Neptune's orbital rotation is almost 165 years, so the planet has not yet completed a full orbit since its discovery. As a result, and because of its extreme distance from the Earth, the dimensions of the planet and its orbit were not well known until 1989, when Voyager 2 became the first – and so far only – spacecraft to visit Neptune.

Neptune's orbit is nearly circular and lies close to the ecliptic. The rotation axis has an Earth-like obliquity of 29.6° and its axial rotation has a period of 16.11 h, which causes a polar flattening of 1.7%. The planet has a radius of 24,766 km and a mean density of 1638 kg m_{-3}. The internal structure of Neptune is probably like that of Uranus: a small rocky core about the size of the Earth is surrounded by a non-layered mixture of rock, water, ammonia, and methane. The atmosphere is predominantly of hydrogen, helium, and methane, which absorbs red light and gives the planet its blue color.

The Voyager 2 mission revealed that Neptune has 13 moons and a faint ring system. The largest of the moons, Triton, has a diameter about 40% of the Earth's and its density (2060 kg m_{-3}) is higher than that of most other large moons in the solar system. Its orbit is steeply inclined at 157° to Neptune's equator, making it the only large natural satellite in the solar system that rotates about its planet in *retrograde* sense. The moon's physical characteristics, which resemble the planet Pluto, and its retrograde orbital motion suggest that Triton was captured from elsewhere in the outer solar system.

1.7 The Outer Solar System: Trans-Neptunian Objects

The solar system extends far beyond the orbit of Neptune. However, at these great distances from the Sun the density of the solar nebula was probably too low for the formation of planets. Since the early 1990s thousands of new objects beyond the orbit of Neptune have been identified. They are known as trans-Neptunian objects (TNOs), defined as any object in orbit around the Sun at a greater average distance than Neptune. The TNOs include Pluto and its moon Charon, as well as numerous other bodies. Most are small, but at least one, Eris, is comparable in size to Pluto. The objects form three groups according to the size of their orbit: the Kuiper belt, the scattered disk, and the Oort Cloud. Their composition is similar to that of comets, i.e., mainly ice, although some have densities high enough to suggest other rock-like components.

1.7.1 Pluto and Charon

Until its reclassification in 2006 as a "dwarf planet," Pluto was the smallest planet in the solar system, about two-thirds the diameter of the Earth's moon. It has many unusual characteristics. Its orbit has the largest inclination to the ecliptic (17.1°) of any planet and it is highly elliptical (eccentricity 0.249), with aphelion at 49.3 AU and perihelion at 29.7 AU. This brings Pluto inside Neptune's orbit for 20 years of its 248-year orbital period; the paths of Pluto and Neptune do not intersect. The orbital period is resonant with that of Neptune in the ratio 3:2 (i.e., Pluto's period is exactly 1.5 times Neptune's). These features preclude any collision between the planets.

Pluto is so far from the Earth that it appears only as a speck of light to Earth-based telescopes and its surface features can be resolved only broadly with the Hubble Space Telescope. It was discovered fortuitously in 1930 after a systematic search for a more distant planet to explain presumed discrepancies in the orbit of Neptune, which, however, were later found to be due to inaccurate estimates of Neptune's mass. The mass and diameter of Pluto were uncertain until in 1978 a moon, Charon, was found to be orbiting Pluto at a distance now measured as 19,570 km. Pluto's mass is only 0.218% that of the Earth. Charon's mass is 12.2% of Pluto's, making it the largest moon in the solar system relative to its primary planet.

Until 2015 Pluto was the only planet that had not been visited by a spacecraft. In 2006 – the same year that Pluto was reclassified as a dwarf planet – the New Horizons spacecraft was launched on a dedicated mission to explore Pluto in a fly-by before venturing out into the Kuiper belt (Section 11.3.2.2). Pluto was reached in 2015 after a flight lasting nine years. The mission was an impressive technical achievement, overcoming difficulties associated with the great distance from Earth to Pluto (4.7 billion km), which is so immense that radio signals take 4.5 h to travel each way. The low power on board the spacecraft makes data transfer so slow that it takes 15 months to complete. New Horizons examined the atmospheres and the surfaces of Pluto and Charon, greatly enhancing existing knowledge of the planet and its main moon.

The radii of Pluto and Charon were found to measure 1188 km and 606 km, respectively, yielding estimated densities of 1854 kg m^{-3} for Pluto and 1702 kg m^{-3} for Charon. Both Pluto and Charon are inferred to have a differentiated internal structure, in which a rocky core with radius of about 70% of the body is surrounded by a water–ice outer shell. One theory of their origin is that they result from the collision of two earlier bodies, the remnants of which have subsequently orbited each other.

Pluto's rotation axis is inclined at about 122° to its orbital plane, so the planet's axial rotation is *retrograde*, and has a period of 6.387 days. Charon also orbits Pluto in a retrograde sense. As a result of tidal forces, Charon's orbital period is synchronous with both its own axial rotation and Pluto's. Thus, the planet and moon constantly present the same faces to each other. The rotational synchronism and the large relative mass of Charon cause Pluto–Charon to resemble a double planet. However, their common barycenter lies outside each body and their different densities suggest that the bodies originated independently.

Pluto has four other moons, each tens of kilometers in diameter, beyond the orbit of Charon. In order of increasing distance beyond Charon the moons are Styx, Nix, Kerberos, and Hydra, which orbits Pluto at a distance of 65,000 km. The orbital planes of all the moons are nearly circular and coplanar with Pluto's equatorial plane, thus forming a plane that is steeply inclined to Pluto's own orbit (Fig. 1.6). As Pluto moves around the Sun the orbital plane of Charon is at times edge-on to the Earth. The New Horizons mission revealed cratering on the surfaces of the moons which suggests that they have a similar ancient age. It is conjectured that they have a common origin in a collision between Pluto and a Kuiper belt body. However, the origins of Pluto, Charon, and the smaller moons are still unknown.

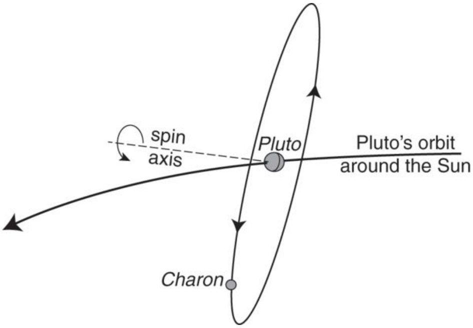

spin
axis
Pluto
Pluto's orbit
around the Sun
Charon

Fig. 1.6 Like Uranus, but in contrast to the other planets, the rotation axis of Pluto has a large obliquity and lies close to the orbital plane of the planet. Charon and the other moons have orbits tilted at less than a degree to Pluto's equator, forming a plane steeply inclined to Pluto's orbit (data source: NASA).

1.7.2 The Kuiper Belt, Scattered Disk Objects, and the Oort Cloud

The *Kuiper belt* extends beyond the mean radius of Neptune’s orbit at 30 AU to a distance of about 50 AU (Fig. 1.7). This disk-shaped region, close to the ecliptic plane, contains thousands of objects in orbit around the Sun. According to some estimates there are more than 35,000 Kuiper belt objects larger than 100 km in diameter, so they are much larger and more numerous than the asteroids. Some have orbital periods that are in resonance with the orbit of Neptune, and this has given rise to some curious appellations for them. Objects like Pluto with orbital periods in 3:2 resonance with Neptune are called plutinos, those further out in the belt with periods in 2:1 resonance are called twotinos, and objects in intermediate orbits are called cubewanos. The Kuiper belt objects are all largely icy in composition, and some of them are quite large. For example, Quaoar, in an orbit with semi-major axis 43.5 AU, has a diameter of 1260 km and so is about the same size as Pluto’s moon, Charon.

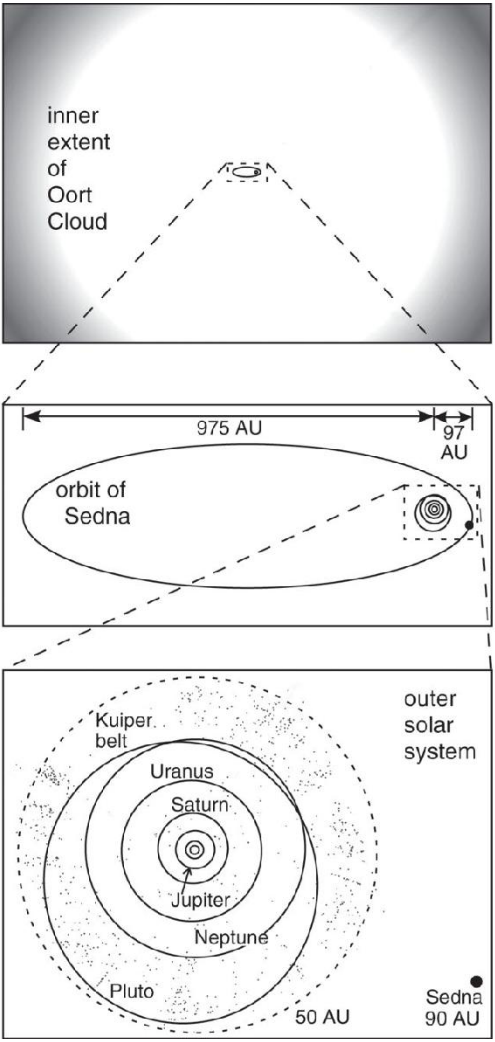

inner
extent
of
Oort
Cloud
975 AU
97
AU
orbit of
Sedna
outer
solar
system
Kuiper
belt
Uranus
Saturn
Jupiter
Neptune
Pluto
50 AU
Sedna
90 AU

Fig. 1.7 The relative sizes of the Oort Cloud and Kuiper belt in relation to the orbits of the outer planets. The inner planets and Sun are contained within the innermost circle of the lower part of the figure (data source: NASA/JPL-Caltech).

After exploring Pluto and Charon, the New Horizons spacecraft continued out into the Kuiper belt. On January 1, 2019 it explored a TNO designated 2014 MU_{69}, passing it at a closest approach of 3500 km. At 43.4 AU from the Sun and 6.5 billion km from the Earth, this is the most distant object yet visited by a spacecraft. First discovered by the Hubble Space Telescope in 2014, the TNO was popularly renamed

Ultima Thule during the fly-by mission to signify its remoteness. It proved to be a binary object, consisting of two joined icy spheres with diameters 17 km and 14 km, and a joined length of 31 km.

Objects in orbit at mean distances greater than 50 AU are called *scattered disk objects*. A large TNO, 2003UB$_{313}$, was identified in 2003 and confirmed in 2005 during a long-term search for distant moving objects in the solar system. On the basis of its reflectivity this object is larger than Pluto, and was at first considered to be the tenth planet in the solar system. It has an orbital period of 557 years, a highly elliptical orbit inclined at 44° to the ecliptic, and is currently near to aphelion. Its present heliocentric distance of 97 AU makes it the most distant known object in the solar system. Now named Eris, it is classified together with Pluto and the asteroid Ceres in the new category of “dwarf planet.”

In 2004 another TNO, Sedna, was discovered at a distance of 90 AU (Fig. 1.7). It is presently closer to the Sun than Eris, but its extremely elliptical orbit (eccentricity 0.855, inclination 12°) takes Sedna further into the outer reaches of the solar system than any known object. Its orbital period is 12,500 years and its aphelion lies at about 975 AU. The object is visible to astronomers only as a tiny speck, so apart from its orbit not much is known about it. It is considered to be the only known object that may have originated in the Oort Cloud.

The *Oort Cloud* is thought to be the source of most new comets that enter the inner solar system. It is visualized as a spherical cloud of icy objects at an enormous distance – between 50,000 and 100,000 AU (roughly one light-year) – from the Sun. The Oort Cloud has never been observed, but its existence has been confirmed from work on cometary orbits. It plays a central role in models of the origin of comets.

1.8 *Suggestions for Further Reading*

Beatty, J. K., Petersen, C. C., and Chaikin, A. (eds.) 1999. *The New Solar System* (4th ed.). Cambridge, MA: Sky Publishing Corp and Cambridge University Press.

Lang, K. R. 2011. *The Cambridge Guide to the Solar System*. Cambridge: Cambridge University Press.

Lowman, P. D. 2002. *Exploring Space, Exploring Earth: New Understanding of the Earth from Space Research*. Cambridge: Cambridge University Press.

Rothery, D. A., McBride, N., and Gilmour, I. (eds.) 2018. *An Introduction to the Solar System*. Cambridge: Cambridge University Press.

1.9 Review Questions

1. What are Kepler's three laws of planetary motion? Which law is a result of the conservation of momentum? Which law is a result of the conservation of energy?

2. If a planet existed in place of the asteroid belt, what would Bode's law predict for the radius of its orbit? What would be the period of its orbital motion around the Sun?

3. What is the nebular hypothesis for the origin of the solar system?

4. Which is the largest planet? How many times larger is it than the Earth? How does its mass compare to that of the Sun?

5. Which is the smallest planet? How do its mass and radius compare to those of the Earth and the Moon?

1.10 Exercises

1. Measured from a position on the Earth's surface at the equator, the angle between the direction to the Moon and a reference direction (distant star) in the plane of the Moon's orbit is $11^\circ\ 57'$ at 8 p.m. one evening and $14^\circ\ 32'$ at 4 a.m. the following morning. Assuming that the Earth, Moon, and reference star are in the same plane, and that the rotation axis is normal to the plane, estimate the approximate distance between the centers of the Earth and Moon.

2. The eccentricity e of the Moon's orbit is 0.0549 and the mean orbital radius r_L is 384,100 km.

(a) Calculate the lengths of the principal axes a and b of the Moon's orbit.

(b) How far is the center of the Earth from the center of the elliptical orbit?

(c) Calculate the distances of the Moon from the Earth at perigee and apogee.

3. If the Moon's disk subtends a maximum angle of 0° 31 min 36.8 s at the surface of the Earth, what is the Moon's radius?

4. Bode's law (Eq. 1.5) gives the mean orbital radius of the nth planet from the Sun (counting the asteroid belt) in astronomical units. It fits the observations well except for Neptune ($n = 9$) and Pluto ($n = 10$). Calculate the orbital radii of Neptune and Pluto predicted by Bode's law, and compare the results with the observed values (Table 1.2). Express the discrepancies as percentages of the predicted distances.

5. An ambulance passes a stationary observer at the side of the road at a speed of 60 km/h. Its dual-tone siren emits alternating tones with frequencies of 700 and 1700 Hz. What are the dual frequencies heard by the observer at 22°C (a) before and (b) after the ambulance passes? [Assume that the speed of sound in air, c, in m s^{-1} at temperature T (°C) is given by $c = 331 + 0.607\ T$.]

6. A spacecraft landing on the Moon uses the Doppler effect on radar signals transmitted at a frequency of 5 GHz to determine the landing speed. The pilot discovers that the precision of the radar instrument has deteriorated to ± 100 Hz. Is this adequate to ensure a safe landing? [Assume speed of light = 300,000 km s^{-1}.]

7. Explain with the aid of a sketch the relationship between the length of a day and the length of a year on the planet Mercury (see Section 1.5.1).

8. The rotations of the planet Pluto and its moon Charon about their own axes are synchronous with the revolution of Charon about Pluto. Show with the aid of simple sketches that Pluto and Charon always present the same faces to each other.

9. The barycenter of a star and its planet – or of a planet and its moon – is the center of mass of the pair. Using the mass and radius of primary body and satellite, and the orbital radius of the satellite, as given in Tables 1.1–1.3, calculate the location of the barycenter of the following pairs of bodies. In each case, does the barycenter lie inside or outside the primary body?

(a) Sun and Earth

(b) Sun and Jupiter

(c) Earth and Moon

(d) Pluto (mass $= 1.303 \times 10^{22}$ kg; radius $= 1188$ km) and Charon (mass $= 1.586 \times 10^{21}$ kg; radius $= 606$ km); the radius of Charon's orbit about Pluto is 19,590 km.

10. A planet with radius R has a mantle with uniform density ρ_m

enclosing a core with radius r_c and uniform density ρ_c. Show that the mean density of the planet ρ is given by

$$\frac{\rho - \rho_m}{\rho_c - \rho_m} = \left(\frac{r_c}{R}\right)^3.$$

11. Assuming that the internal structure of the Moon is a set of concentric shells with different density (Fig. 1.5) as simplified in the following table, compute the Moon's mean density.

Internal layer	**Radius (km)**	**Density (kg m^{-3})**
Solid inner core	0–240	8000
Fluid outer core	240–330	5000
Mantle	330–1738	3300

12. The moment of inertia I of a spherical shell of density ρ and outer and inner radii R_2 and R_1, respectively, about its axis of symmetry is given by the formula:

$$I = \frac{8\pi}{15}\rho\left(R_2^5 - R_1^5\right)$$

(a) Assuming the internal structure of the Moon from the previous exercise, compute the Moon's moment of inertia about its rotation axis.

(b) The moment of inertia of a sphere can be written $I = kMR^2$, where M is its mass and R its radius. Calculate the value of k for the Moon.

13. Using the data in Table 1.2, show that the planets conform to Kepler's third law. What is the meaning of the constant in your result?

2

Plate Tectonics

Preview

The lithosphere – the thin outer shell of the Earth – is stronger than the underlying mantle. Geodynamic activity driven by heat in the planet’s interior has caused it to subdivide into a number of thin plates, several hundred to several thousand kilometers in horizontal extent. They are in constant motion at speeds of a few centimeters per year. The relative motion at plate margins where they adjoin results in tectonic activity, characterized by earthquakes and volcanism. We distinguish three types of margin – spreading centers, subduction zones, and transform faults – and describe the seismic, gravity, and magnetic data that characterize them. We explain how reconstructions of plate positions in the geological past are obtained and what they tell us about the planet’s geodynamic history.

2.1 Historical Introduction

The Earth is a dynamic planet, perpetually changing both externally and internally. Its surface is constantly being altered by *endogenic* processes (i.e., of *internal* origin) resulting in volcanism and tectonism, as well as by *exogenic* processes (i.e., of *external* origin) such as erosion and deposition. These processes have been active throughout geological history. Volcanic explosions like the 1980 eruption of Mt. St. Helens in the northwestern United States can transform the surrounding landscape virtually instantaneously. Earthquakes also cause sudden changes in the landscape, sometimes producing faults with displacements of several meters in seconds. Weather-related erosion of surface features occasionally occurs at dramatic rates, especially if rivers overflow or landslides are triggered. The Earth's surface is also being changed constantly by less spectacular geological processes at rates that are extremely slow in human terms. Regions that have been depressed by the loads of past ice-sheets are still rebounding vertically at rates of up to several millimeters per year. Tectonic forces cause mountains to rise at similar uplift rates, while the long-term average effects of erosion on a regional scale may occur at rates of centimeters per year or more. Where human activity affects the landscape, erosion rates can be an order of magnitude higher. On a larger scale the continents move relative to each other at speeds of up to several centimeters per year for time intervals lasting millions of years. Extremely long times are represented in geological processes. This is evident in the lengths of the time intervals in a geological timescale (see Fig. 8.2).

The Earth's interior is also in motion. The mantle appears hard and solid to seismic waves, but is believed to exhibit a softer, viscous behavior over long geological time intervals, flowing (or "creeping") at rates of several centimeters per year. Deeper inside the Earth, the liquid core probably flows at a geologically rapid rate of a few tenths of a millimeter per second.

Geologists have long been aware of the Earth's dynamic condition. Several hypotheses have attempted to explain the underlying mechanisms. In the late nineteenth and early twentieth centuries geological orthodoxy favored the hypothesis of a contracting Earth. Mountain ranges were thought to have formed on its shrinking surface like wrinkles on a desiccating apple. Horizontal tectonic displacements were known, but were considered to be a by-product of more important vertical motions. The realization that large overthrusts played an important role in the formation of nappe structures in the Alps implied amounts of horizontal shortening that were difficult to accommodate in the contraction hypothesis. A new school of thought emerged in which mountain-building was depicted as a consequence of horizontal displacements.

A key observation in this context was the congruity between the opposing coasts of the South Atlantic, especially the similar shapes of the coastlines of Brazil and Africa. As early as 1620, despite the inaccuracy and incompleteness of early seventeenth-century maps, Francis Bacon drew attention to the parallelism of the Atlantic-bordering coastlines. In 1858 Antonio Snider constructed a map showing relative movements of the circum-Atlantic continents, although he did not maintain the shapes of the coastlines. In the late nineteenth century the Austrian geologist Eduard Suess coined the name *Gondwana* for a proposed great southern continent that existed during late Paleozoic times. It embodied Africa, Antarctica, Arabia, Australia, India, and South America, and lay predominantly in the southern hemisphere. The Gondwana continents are now individual entities and some (e.g., India, Arabia) no longer lie in the southern hemisphere, but they are often still called the "southern continents." In the Early Paleozoic, the "northern continents" of North America (including Greenland), Europe, and most of Asia also formed a single continent, called *Laurussia*. It joined with Gondwana and Siberia in the late Paleozoic to form a single supercontinent, called *Pangea* (or

Pangaea), which subsequently existed for about 100 million years (Myr). Implicit in these reconstructions is the idea that the continents subsequently reached their present positions by slow horizontal displacements across the surface of the globe.

2.2 Continental Drift

The "displacement hypothesis" of continental movements matured early in the twentieth century. In 1908 F. B. Taylor related the world's major fold-belts to convergence of the continents as they moved away from the poles, and in 1911 H. B. Baker reassembled the Atlantic-bordering continents together with Australia and Antarctica into a single continent; regrettably he omitted Asia and the Pacific. However, the most vigorous proponent of the displacement hypothesis was Alfred Wegener, a German meteorologist and geologist. In 1912 Wegener suggested that all of the continents were together in the Late Paleozoic, so that the land area of the Earth formed a single landmass (Fig. 2.1). He coined the name *Pangea* (Greek for "all Earth") for this supercontinent, which he envisioned was surrounded by a single ocean (*Panthalassa*). Wegener referred to the large-scale horizontal displacement of crustal blocks having continental dimensions as *Kontinentalverschiebung*. The anglicized form, *continental drift*, implies additionally that displacements of the blocks take place slowly over long time intervals.

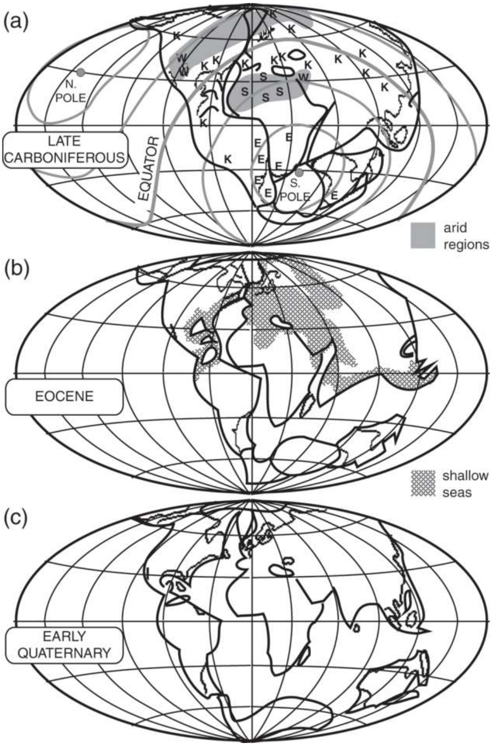

(a)
N. POLE
LATE CARBONIFEROUS
EQUATOR
S. POLE
K
S
W
E
arid regions
(b)
EOCENE
shallow seas
(c)
EARLY QUATERNARY

Fig. 2.1 (a) Wegener's reconstruction of Pangea in the Late Carboniferous, showing estimated positions of the North and South poles and paleo-equator. *Shaded areas*: locations of shallow seas; K: coal deposits; S: salt deposits; W: desert regions; E: ice-sheets (modified after Köppen and Wegener, 1924). Relative positions of the continents are shown in (b) the Eocene (*shaded areas*: shallow seas) and (c) the Early Quaternary (after Wegener, 1922). The latitudes and longitudes are arbitrary.

2.2.1 Pangea

As a meteorologist, Wegener was especially interested in paleoclimatology. For the first half of the twentieth century the best evidence for the continental drift hypothesis and the earlier existence of Pangea consisted of geological indicators of earlier paleoclimates. In particular, Wegener observed a much better alignment of regions of Permo-Carboniferous glaciation in the southern hemisphere when the continents were in the reconstructed positions for Gondwana instead of their present positions. His reconstruction of Pangea brought Carboniferous coal deposits into alignment and suggested that the positions of the continents relative to the Paleozoic equator were quite different from their modern ones. Together with W. Köppen, a fellow German meteorologist, he assembled paleoclimatic data that showed the distributions of coal deposits (evidence of moist temperate zones), salt, gypsum, and desert sandstones (evidence of dry climate) for several geological eras (Carboniferous, Permian, Eocene, Quaternary). When plotted on Wegener's reconstruction maps, the paleoclimatic data for each era formed climatic belts just like today; namely, an equatorial tropical rain belt, two adjacent dry belts, two temperate rain belts, and two polar ice caps (Fig. 2.1a).

Wegener's continental drift hypothesis was bolstered in 1937 by the studies of a South African geologist, Alexander du Toit, who noted sedimentological, paleontological, paleoclimatic, and tectonic similarities between western Africa and eastern South America. These favored the Gondwana reconstruction rather than the present configuration of continents during the Late Paleozoic and Early Mesozoic.

Some of Wegener's theories were largely conjectural. On the one hand, he reasoned correctly that the ocean basins are not permanent. Yet he envisioned the sub-crustal material as capable of viscous yield over long periods of time, enabling the continents to drift through the ocean crust like ships through water. This model met with profound skepticism among geologists. He believed, in the face of strong opposition from physicists, that the Earth's geographic axis had moved with time, instead of the crust moving relative to the fixed poles. His timing of the opening of the Atlantic (Fig. 2.1b, c) was faulty, requiring a large part of the separation of South America from Africa to take place since the Early Pleistocene (i.e., in the past two million years or so). Moreover, he was unable to offer a satisfactory driving mechanism for continental drift. His detractors used the disprovable speculations to discredit his better-documented arguments in favor of continental drift.

2.2.2 Computer-Assisted Reconstructions

Wegener pointed out that it was not possible to fit the continents together using their present coastlines, which are influenced by recent sedimentary deposits at the mouths of major rivers as well as the effects of coastal erosion. The large areas of continental shelf must also be taken into account, so Wegener matched the continents at about the edges of the continental shelves, where the continental slopes plunge into the oceanic basins. The matching was visual and inexact by modern standards, but more precise methods only became available in the 1960s with the development of electronic computers.

In 1965 E. C. Bullard, J. E. Everett, and A. G. Smith used a computer to match the relative positions of the continents bounding the Atlantic Ocean (Fig. 2.2). They digitized the continental outlines at approximately 50 km intervals for different depth contours on the continental slopes, and selected the fit of the 500 fathom (900 m) depth contour as optimum. The traces of opposite continental margins were matched by an iterative procedure. One trace was rotated relative to the other (about a pole of relative rotation) until the differences between the traces were minimized; the procedure was then repeated with different rotation poles until the best fit was obtained. The optimum fit is not perfect, but has some overlaps and gaps. Nevertheless, the analysis gives an excellent geometric fit of the opposing coastlines of the Atlantic.

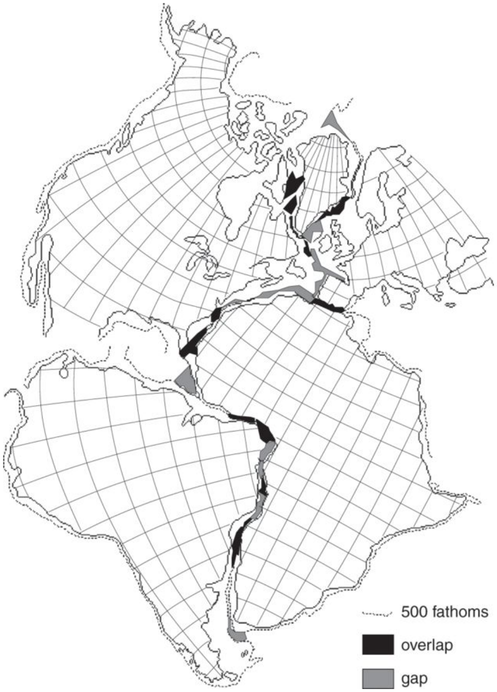


Fig. 2.2 Computer-assisted fit of the Atlantic-bordering continents at the 500 fathom (900 m) depth.

Source: E.C. Bullard, J.E. Everett, and A.G. Smith, A Symposium on Continental Drift, IV. The fit of the continents around the Atlantic, *Phil. Trans. R. Soc. London, Ser. A* 258, 41–51, 1965.

A few years later, A. G. Smith and A. Hallam used the same computer-assisted technique to match the coastlines of the southern continents, also at the 500 fathom depth contour (Fig. 2.3). They obtained an optimum geometric reconstruction of Gondwana similar to the visual match suggested by du Toit in 1937; it probably represents the geometry of Gondwana that existed in the Late Paleozoic and Early Mesozoic. It is not the only possible good geometric fit, but it also satisfies other geological evidence. At various times in the Jurassic and Cretaceous, extensional plate margins formed within Gondwana, causing it to subdivide to form the present "southern continents." Their dispersal to their present positions took place largely in the Late Cretaceous and Tertiary.

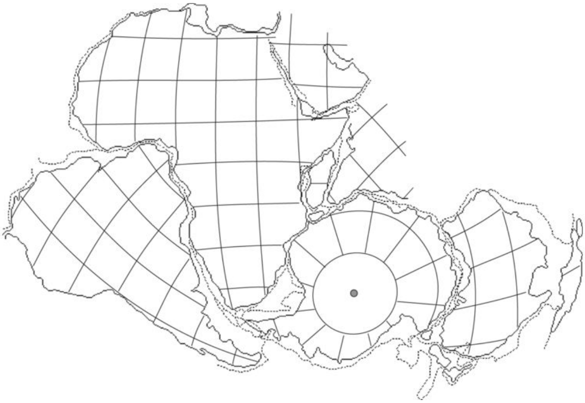


Fig. 2.3 Computer-assisted fit of the continents that form Gondwana.

Source: A.G. Smith and A. Hallam, The fit of the southern continents, *Nature* 225, 139–144, 1970.

Pangea existed only in the Late Paleozoic and Early Mesozoic. Geological and geophysical evidence argues in favor of the existence of its northern and southern constituents – Laurussia and Gondwana – as separate entities in the Early Paleozoic and Precambrian. An important source of data bearing on continental reconstructions in ancient times and the drift of the continents is provided by *paleomagnetism*, which is the record of the Earth's ancient magnetic field. Paleomagnetism is described in Chapter 12 and summarized in the next section.

2.2.3 Paleomagnetism and Continental Drift

In the late nineteenth century geologists discovered that rocks can carry a stable record of the geomagnetic field direction at the time of their formation. From the magnetization direction it is possible to calculate the position of the magnetic pole at that time; this is called the virtual geomagnetic pole (VGP) position. Averaged over a time interval longer than a few tens of thousands of years, the mean VGP position coincides with the geographic pole, as if the axis of the mean geomagnetic dipole field were aligned with the Earth's rotation axis. This correspondence can be proved for the present geomagnetic field, and a fundamental assumption of paleomagnetism – called the "geocentric axial dipole hypothesis" – is that it has always been valid. The hypothesis can be verified for rocks and sediments up to a few million years old, but its validity has to be assumed for earlier geological epochs. However, the self-consistency of paleomagnetic data and their compatibility with continental reconstructions argue that the axial dipole hypothesis is also applicable to the Earth's ancient magnetic field.

For a particular continent, rocks of different ages give different mean VGP positions. The appearance that the pole has shifted with time is called apparent polar wander (APW). By connecting mean VGP positions of different ages for sites on the same continent, a line is obtained, called the APW path of the continent. Each continent yields a different APW path (Fig. 2.4), which consequently cannot be the record of movement of the pole. Rather, each APW path represents the movement of the continent relative to the pole. By comparing APW paths, the movements of the continents relative to each other can be reconstructed. The APW paths provide strong supporting evidence for continental drift.

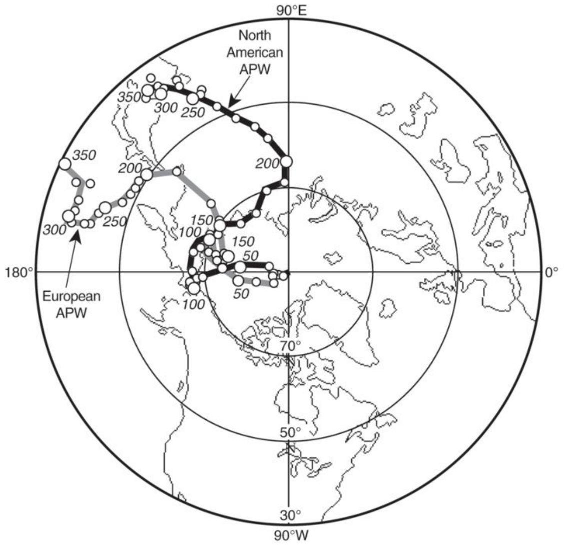


Fig. 2.4 Average apparent polar wander paths for North America and Europe in the past 350 Myr. Numbers on paths denote age in Myr.

Source: E. Irving, Drift of the major continental blocks since the Devonian, *Nature* 270, 304–309, 1977.

Paleomagnetism developed as a geological discipline in the 1950s and 1960s. The first results indicating large-scale continental movement were greeted with some skepticism. In 1956, S. K. Runcorn demonstrated that the paleomagnetic data from Permian and Triassic rocks in North America and Great Britain agreed better if the Atlantic Ocean were closed, i.e., as in the Laurasia configuration. In 1957, E. Irving showed that Mesozoic paleomagnetic data from the “southern continents” were more concordant with du Toit’s Gondwana reconstruction than with the present arrangement of the continents. Since these pioneering studies, numerous paleomagnetic investigations have established APW paths for the different continents. The quality of the paleomagnetic record is good for most geological epochs since the Early Paleozoic. It is the most important geophysical method for reconstructing the pre-Jurassic configuration of the continents.

The consistency of paleomagnetic results leaves little room for doubt that the continents have changed position relative to each other throughout geological time. This lends justification to the concept of continental drift, but it does not account for the mechanism by which it has taken place. We now know that continental drift is a consequence of the motions of the lithospheric plates on which the continents are located (Section 2.3.1). Another aspect of the paleomagnetic record – the history of magnetic field polarity rather than the APW paths – has played a key role in deducing the mechanism. The explanation requires an understanding of the Earth's internal structure, the distribution of seismicity and the importance of the ocean basins.

2.3 Earth Structure

Early in the twentieth century it became evident from the study of seismic waves that the interior of the Earth has a radially layered structure, like that of an onion (Fig. 2.5). The boundaries between the layers are marked by abrupt changes in seismic velocity or velocity gradient. Each layer is characterized by a specific set of physical properties determined by the composition, pressure, and temperature in the layer. The four main layers are the crust, mantle, and the outer and inner cores. Their properties are described in detail in Section 7.4.4 and summarized briefly here.

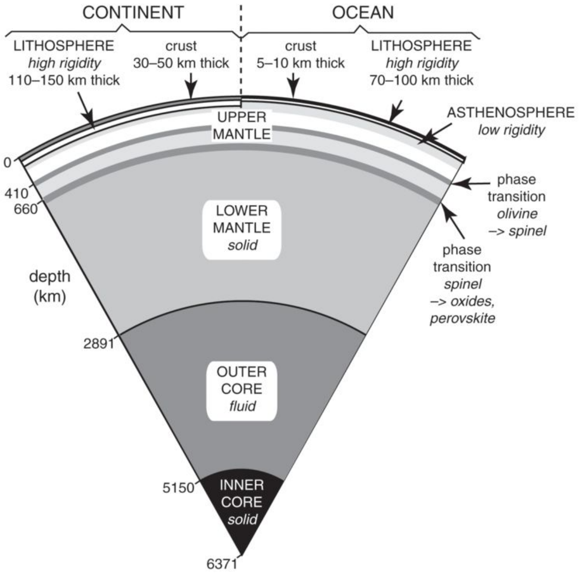

CONTINENT
OCEAN
LITHOSPHERE
high rigidity
110–150 km thick
crust
30–50 km thick
crust
5–10 km thick
LITHOSPHERE
high rigidity
70–100 km thick
ASTHENOSPHERE
low rigidity
UPPER MANTLE
0
410
660
phase transition
olivine
–> spinel
LOWER MANTLE
solid
phase transition
spinel
–> oxides, perovskite
depth (km)
2891
OUTER CORE
fluid
5150
INNER CORE
solid
6371

Fig. 2.5 Simplified layered structure of the Earth's interior showing the depths of the most important seismic discontinuities.

At depths of a few tens of kilometers under continents and less than 10 km beneath the oceans seismic velocities increase sharply. This seismic discontinuity, discovered in 1909 by A. Mohorovičić, represents the boundary between the *crust* and *mantle*. R. D. Oldham noted in 1906 that the travel times of seismic compressional waves that traversed the body of the Earth were greater than expected; the delay was attributed to a *fluid outer core*. Support for this idea came in 1914, when B. Gutenberg described a shadow zone for seismic waves at epicentral distances greater than about 105°. Just as light waves cast a shadow of an opaque object, seismic waves from an earthquake cast a shadow of the core on the opposite side of the world. Compressional waves can in fact pass through the liquid core. They appear, delayed in time, at epicentral distances larger than 143°. In 1936 I. Lehmann observed the weak arrivals of compressional waves in the gap between 105° and 143°. They are interpreted as evidence for a *solid inner core*.

2.3.1 Lithospheric Plates

The radially layered model of the Earth's interior assumes spherical symmetry, which is not quite true. For example, the crust and upper mantle show important lateral variations. The crust and uppermost mantle down to a depth of about 70–100 km under deep ocean basins and 100–150 km under continents are rigid, forming a hard outer shell called the *lithosphere*. Beneath the lithosphere lies the *asthenosphere*, a layer in which seismic velocities often decrease, suggesting lower rigidity. It is about 150 km thick, although its upper and lower boundaries are not sharply defined. This weaker layer is thought to be partially molten; it is able to flow over long periods of time like a viscous liquid or plastic solid, in a way that depends on temperature and composition. The asthenosphere plays an important role in plate tectonics, because it makes possible the relative motions of the overlying lithospheric plates.

The brittle condition of the lithosphere causes it to fracture when strongly stressed. The rupture produces an earthquake, which is the violent release of elastic energy due to sudden displacement on a fault plane. Earthquakes are not distributed evenly over the surface of the globe, but occur predominantly in well-defined narrow seismic zones that are often associated with volcanic activity (Fig. 2.6). These are: (a) the circum-Pacific "ring of fire"; (b) a sinuous belt running from the Azores through North Africa and the Alpine–Dinaride–Himalayan mountain chain as far as Southeast Asia; and (c) the world-circling system of oceanic ridges and rises.

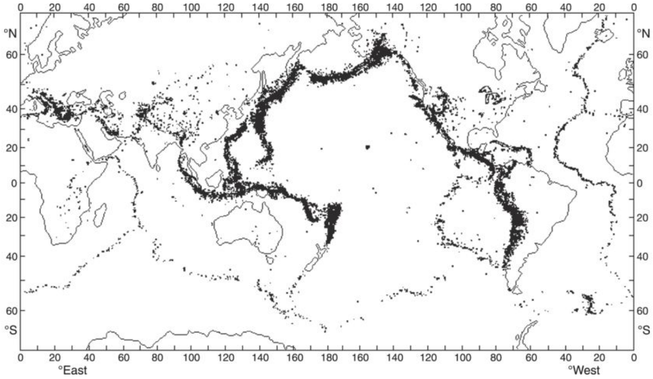


Fig. 2.6 The geographical distribution of epicenters of 30,000 earthquakes for the years 1961–1967 illustrates the tectonically active regions of the Earth.

Source: M. Barazangi and J. Dorman, World seismicity maps compiled from ESSA, Coast and Geodetic Survey, epicenter data, 1961–1967, *Bull. Seism. Soc. Amer.* 59, 369–380, 1969.

The seismic zones subdivide the lithosphere laterally into *tectonic plates* (Fig. 2.7). Twelve major plates were recognized (Antarctica, Africa, Eurasia, India, Australia, Arabia, Philippines, North America, South America, Pacific, Nazca, and Cocos). The African plate was later divided into a western plate (called Nubia) and an eastern plate (called Somalia) at the East African Rift Zone, which is considered to be an incipient spreading margin. In addition there are several minor plates (e.g., Scotia, Caribbean, Juan de Fuca) and many small plates, which are often called microplates (e.g., the Sandwich microplate in the South Atlantic). Models of current plate motions, based on the interpretation of magnetic anomaly patterns at ocean ridges and first-motion directions in earthquakes, show rates of separation at plate boundaries that range from ~20 mm yr^{-1} in the North Atlantic to ~140 mm yr^{-1} on the East Pacific Rise (Fig. 2.7). Rates of closure are observed to range from ~10 mm yr^{-1} between Africa and Eurasia to ~70 mm yr^{-1} between the Nazca plate and South America.

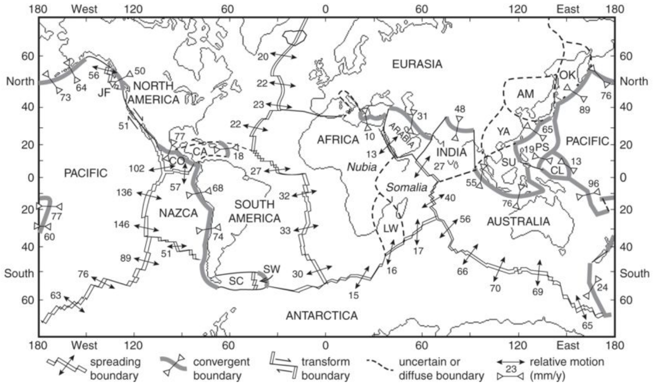

180
West
120
60
0
60
120
East
180
60
North
40
20
0
20
40
South
60
EURASIA
NORTH AMERICA
JF
PACIFIC
CO
CA
NAZCA
SOUTH AMERICA
AFRICA
Nubia
Somalia
ARABIA
INDIA
LW
AM
OK
YA
SU
PS
CL
AUSTRALIA
SW
SC
ANTARCTICA
50
56
64
73
51
77
18
102
136
57
68
146
74
89
51
76
63
77
60
20
22
23
22
27
32
33
30
15
16
17
10
13
31
48
27
40
56
66
55
76
70
69
65
19
13
89
76
96
24
65
spreading boundary
convergent boundary
transform boundary
uncertain or diffuse boundary
23
relative motion (mm/y)

Fig. 2.7 Major and minor lithospheric plates. The arrows indicate relative velocities in mm yr^{-1} at active plate margins, calculated from the MORVEL model of current plate motions. Identification of small plates: AM, Amur; CA, Caribbean; CL, Caroline; CO, Cocos; JF, Juan de Fuca; LW, Lwandle; OK, Okhotsk; PS, Philippine Sea; SC, Scotia; SU, Sunda; SW, Sandwich; YA, Yangtze. Despite being extended over tens to hundreds of kilometers, diffuse boundaries, for example between the Australian and Indian plates, are depicted as thin lines in this figure.

Data source: C. DeMets, R.G. Gordon, and D.F. Argus, Geologically current plate motions, *Geophys. J. Int.* 181, 1–80, 2010.

Space geodetic observations are able to measure current strain rates and can identify incipient as well as active plate boundaries. Some uncertain plate boundaries were resolved by this means and previously uncertain boundaries were clarified. For example, a plate margin within the Indo-Australian plate was defined, distinguishing separate Indian and Australian plates that have slightly different rates of rapid northward motion. The boundary between the North and South American plates is vaguely defined, but the plates have clearly distinct rotation poles.

2.4 Types of Plate Margin

An important factor in the evolution of modern plate tectonic theory was the development of oceanography in the years following World War II, when technology designed for warfare was turned to peaceful purposes. The bathymetry of the oceans was charted extensively by echo-sounding, and within a few years several striking features became evident. Deep trenches, more than twice the depth of the ocean basins, were discovered close to island arcs and some continental margins; the Marianas trench is more than 11 km deep. A prominent submarine mountain chain – called an oceanic ridge – was found in each ocean. The oceanic ridges rise to as much as 3000 m above the adjacent basins and form a continuous system, more than 60,000 km in length, that girdles the globe. Unlike continental mountain belts, which are usually less than several hundred kilometers across, the oceanic ridges are 2000–4000 km in width. The ridge system is offset at intervals by long horizontal faults forming fracture zones. These three features – trenches, ridges, and fracture zones – originate from different plate tectonic processes.

The lithospheric plates are very thin in comparison to their breadth (compare Figs. 2.5 and 2.7). Most earthquakes occur at plate margins, and are associated with interactions between plates. Apart from rare intraplate earthquakes, which can be as large and disastrous as the earthquakes at plate boundaries, the plate interiors are aseismic. This suggests that the plates behave rigidly. Analysis of earthquakes allows the direction of displacement to be determined and permits interpretation of the relative motions between plates.

There are three types of plate margin, distinguished by different tectonic processes (Fig. 2.8). The world-wide pattern of earthquakes shows that the plates are presently moving apart at oceanic ridges. Magnetic evidence, discussed below, confirms that the separation has been going on for millions of years. New lithosphere is being formed at these *spreading centers*, so the ridges can be regarded as *constructive* plate margins. The seismic zones related to deep-sea trenches, island arcs, and mountain belts mark places where lithospheric plates are converging. One plate is forced under another there in a so-called *subduction zone*. Because it is thin relative to its breadth, the lower plate bends sharply before descending to depths of several hundred kilometers, where it is absorbed. The subduction zone marks a *destructive* plate margin.

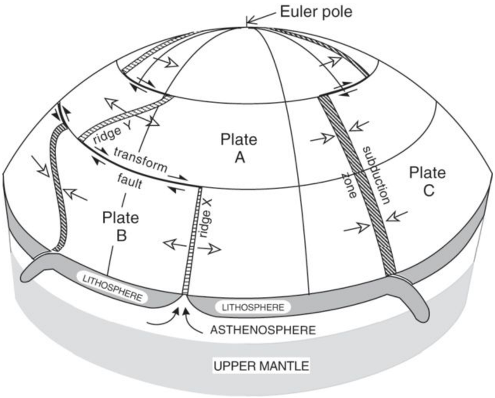


Fig. 2.8 Diagram illustrating the three types of plate margin. Lightly hachured areas symbolize spreading ridges (constructive margins); darker shaded areas denote subduction zones (destructive margins); dark lines mark transform faults (conservative margins). The figure is drawn relative to the pole of relative motion between plates A and B. Small arrows denote relative motion on transform faults; large arrows show directions of plate motion, which can be oblique to the strike of ridge segments or subduction zones.

Constructive and destructive plate margins may consist of many segments linked by horizontal faults. A crucial step in the development of plate tectonic theory was made in 1965 by a Canadian geologist, J. Tuzo Wilson, who recognized that these faults are not conventional transcurrent faults. They belong to a new class of faults, which Wilson called *transform faults.* The relative motion on a transform fault is opposite to what might be inferred from the offsets of bordering ridge segments. At the point where a transform fault meets an oceanic ridge, it transforms the spreading on the ridge to horizontal shear on the fault. Likewise, where such a fault meets a destructive plate margin it transforms subduction to horizontal shear.

The transform faults form a *conservative* plate margin, where lithosphere is neither created nor destroyed; the boundary separates plates that move past each other horizontally. This interpretation was documented in 1967 by L. Sykes, an American seismologist. He showed that earthquake activity on an oceanic ridge system was confined almost entirely to the transform fault between ridge crests, where the neighboring plates rub past each other. Most importantly, Sykes found that the mechanisms of earthquakes on the transform faults agreed with the predicted sense of strike–slip motion.

Transform faults play a key role in determining plate motions. Spreading and subduction are often assumed to be perpendicular to the strike of a ridge or trench, as is the case for ridge X in Fig. 2.8. This is not necessarily the case. Oblique motion with a component along strike is possible at each of these margins, as on ridge Y. However, because lithosphere is neither created nor destroyed at a conservative margin, the relative motion between adjacent plates must be parallel to the strike of a shared transform fault. Independent pioneering studies by D. P. McKenzie and R. L. Parker (1967) and W. J. Morgan (1968) showed how transform faults could be used to locate the Euler pole of rotation for two plates (see Section 2.9). Using this method, X. Le Pichon in 1968 determined the present relative motions of the major tectonic plates. In addition, he derived the history of plate motions in the geological past by incorporating newly available magnetic results from the ocean basins.

2.5 Sea-Floor Spreading

One of the principal stumbling blocks of continental drift was the inability to explain the mechanism by which drift took place. Wegener had invoked forces related to gravity and the Earth's rotation, which were demonstrably much too weak to drive the continents through the resistant basaltic crust. In 1944 A. Holmes proposed a model that closely resembles the accepted plate tectonic model (see Holmes, 1965). He noted that it would be necessary to remove basaltic rocks continuously out of the path of an advancing continent, and suggested that this took place at the ocean deeps where heavy eclogite "roots" would sink into the mantle and melt. The model envisaged that convection currents in the upper mantle would return the basaltic magma to the continents as plateau basalts, and to the oceans through innumerable fissures. Holmes saw generation of new oceanic crust as a process that was dispersed throughout an ocean basin. At the time of his proposal the existence of the system of oceanic ridges and rises was not yet known.

The important role of oceanic ridges was first recognized by H. Hess in 1962. He suggested that new oceanic crust is generated from upwelling hot mantle material at the ridges, where convection currents in the upper mantle rise to the surface and subsequently spread out laterally. The continents ride on the spreading mantle material, carried along passively by the convection currents. In 1961 R. Dietz coined the expression "sea-floor spreading" for the ridge process. It results in the generation of lineated marine magnetic anomalies at the ridges, which record the history of geomagnetic polarity reversals. Study of these magnetic effects led to the verification of sea-floor spreading.

2.5.1 The Vine–Matthews–Morley Hypothesis

Paleomagnetic studies in the late 1950s and early 1960s of radiometrically dated continental lavas showed that the geomagnetic field has changed polarity at irregular time intervals. For tens of thousands to millions of years the polarity might be normal (as at present), then unaccountably the poles reverse within a few thousand years, so that the north magnetic pole is near the south geographic pole and the south magnetic pole is near the north geographic pole. This state may persist for a long interval before the polarity switches again. The ages of the reversals in the last five million years have been obtained radiometrically, giving an irregular but dated polarity sequence.

A magnetic anomaly is a departure from the theoretical magnetic field at a given location. If the field is stronger than expected, the anomaly is positive; if it is weaker than expected, the anomaly is negative. In the late 1950s magnetic surveys over the oceans revealed remarkable striped patterns of alternately positive and negative magnetic anomalies over large areas of oceanic crust (Fig. 2.9), for which conventional methods of interpretation gave no satisfactory account. In 1963 the English geophysicists F. J. Vine and D. H. Matthews, and, independently, the Canadian geologist L. W. Morley, formulated a landmark hypothesis that explains the origin of the oceanic magnetic anomaly patterns (see also Section 11.4.6).

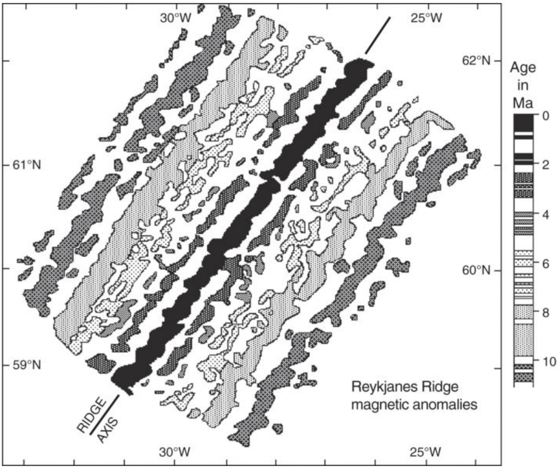


Fig. 2.9 Symmetric striped pattern of magnetic anomalies on the Reykjanes segment of the Mid-Atlantic Ridge southwest of Iceland. The positive anomalies are shaded according to their age, as indicated in the vertical column.

Source: J.R. Heirtzler, X. Le Pichon, and J.G. Baron, Magnetic anomalies over the Reykjanes Ridge, *Deep Sea Res.* 13, 427–443., 1966.

Observations on dredged samples had shown that basalts in the uppermost oceanic crust carry a strong remanent magnetization (i.e., they are permanently magnetized, like a magnet). The Vine–Matthews–Morley hypothesis integrates this result with the newly acquired knowledge of geomagnetic polarity reversals and the Hess–Dietz concept of sea-floor spreading (Fig. 2.10). The basaltic lava is extruded in a molten state. When it solidifies and its temperature cools below the Curie temperature of its magnetic minerals, the basalt becomes strongly magnetized in the direction of the Earth's magnetic field at that time. Along an active spreading ridge, long thin strips of magnetized basaltic crust form symmetrically on opposite sides of the spreading center, each carrying the magnetic imprint of the field in which it formed. Sea-floor spreading can persist for many millions of years at an oceanic ridge. During this time the magnetic field changes polarity many times, forming strips of oceanic crust that are magnetized alternately parallel and opposite to the ambient field, giving the observed patterns of positive and negative anomalies. Thus, the basaltic layer acts like a magnetic tape recorder, preserving a record of the changing geomagnetic field polarity.

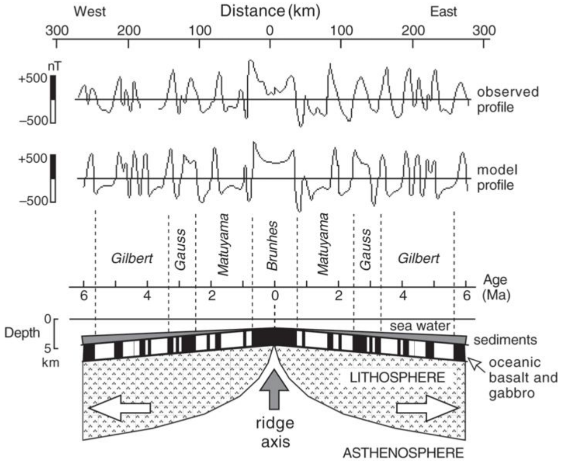

West
Distance (km)
East
300
200
100
0
100
200
300
nT
+500
−500
observed profile
+500
−500
model profile
Gilbert
Gauss
Matuyama
Brunhes
Matuyama
Gauss
Gilbert
6
4
2
0
2
4
6
Age (Ma)
Depth
0
5
km
sea water
sediments
oceanic basalt and gabbro
LITHOSPHERE
ridge axis
ASTHENOSPHERE

Fig. 2.10 Upper: Observed and computed marine magnetic anomalies, in nanotesla (nT), across the Pacific–Antarctica ridge; lower: their interpreted origin in terms of the Vine–Matthews–Morley hypothesis.

Source: W.C. Pitman III and J.R. Heirtzler, Magnetic anomalies over the Pacific–Antarctic ridge, *Science* 154, 1164–1171, 1966.

2.5.2 Rates of Sea-Floor Spreading

The width of a magnetic lineation (or stripe) depends on two factors: the speed with which the oceanic crust moves away from a spreading center, and the length of time that geomagnetic polarity is constantly normal or reversed. The distance between the edges of magnetized crustal stripes can be measured from magnetic surveys at the ocean surface, while the ages of the reversals can be obtained by correlating the oceanic magnetic record with the radiometrically dated reversal sequence determined in subaerial lavas for about the last 4 Myr. When the distance of a given polarity reversal from the spreading axis is plotted against the age of the reversal, a nearly linear relationship is obtained (Fig. 2.11). The slope of the best-fitting straight line gives the average *half-rate of spreading* at the ridge. These are of the order of 10 mm yr^{-1} in the North Atlantic Ocean and 40–140 mm yr^{-1} in the Pacific Ocean. The calculation applies to the rate of motion of crust on one side of the ridge only. In most cases spreading has been symmetric on each side of the ridge (i.e., the opposite sides are moving away from the ridge at equal speeds), so the full rate of separation at a ridge axis is double the calculated half-rate of spreading (Fig. 2.11).

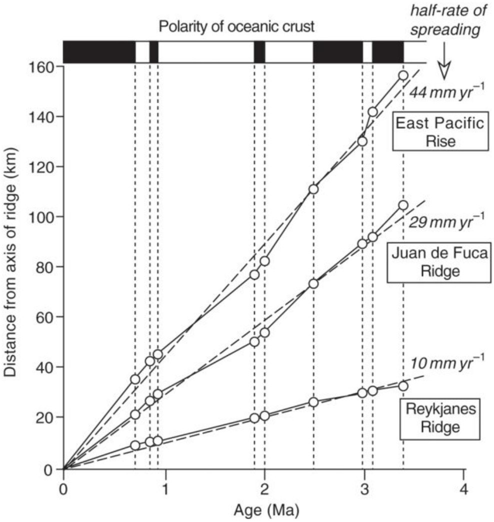


Fig. 2.11 Computation of half-rates of sea-floor spreading at

different spreading centers by measuring the distances to anomalies with known radiometric ages.

Source: F.J. Vine, Spreading of the ocean floor: new evidence, *Science* 154, 1405–1415, 1966.

Knowing the spreading rates at ocean ridges makes it possible to date the ocean floor. The direct correlation between polarity sequences measured in continental lavas and derived from oceanic anomalies is only possible for the last 4 Myr or so. Close to the axial zone, where linear spreading rates are observed (Fig. 2.11), magnetic anomalies can be dated by simple extrapolation, converting the striped pattern into an age map (Fig. 2.9). Detailed magnetic surveying has revealed that most of the ocean floor is characterized by lineated magnetic anomalies. A continuous sequence of anomalies since the Late Cretaceous is preceded by an interval in which no reversals occurred; this Quiet Interval was itself preceded by an earlier Mesozoic reversal sequence. Away from ridge systems the magnetic anomalies cannot be dated accurately enough by simple extrapolation of the linear rate near the ridge. For older parts of the oceans the development of magnetostratigraphy in sedimentary rocks (Section 12.3.2) enabled the identification, correlation, and dating of key magnetic anomalies.

The oceanic anomaly polarity sequence has been converted to a magnetic polarity timescale in which each polarity reversal is accorded an age. In turn, this allows the pattern of magnetic anomalies in the ocean basins to be converted to a map of the age of the ocean basins (as in Fig. 12.35). The oldest areas of the oceans lie close to northwest Africa and eastern North America, as well as in the northwest Pacific. These areas formed during the early stages of the breakup of Pangea in the Early Jurassic. The ages of the ocean basins have been confirmed by drilling through the sediment layers that cover the ocean floor and into the underlying basalt layer. Beginning in the late 1960s and extending until the present, this immensely expensive undertaking has been carried out in the Deep Sea Drilling Project (DSDP) and its successor, the Ocean Drilling Project (ODP). These multinational projects, under the leadership of the USA, are prime examples of open scientific cooperation on an international scale.

2.5.3 Models of Plate Motions

The rates of current plate motion determined from axial anomaly patterns are average values over several million years. Modern geodetic methods (Section 3.5) allow these rates to be tested directly. The satellite-based Global Positioning System (GPS), in use since the 1980s, has made it possible to determine the location of a point on the Earth with a horizontal accuracy of about 5 m in civilian use; in conjunction with an augmentation service the accuracy improves to a few centimeters. Deployed in scientific networks, or when observations at a given GPS receiver are carried out over long periods of time, the accuracy of location improves dramatically to the millimeter level. It has thus been possible to use GPS directly to observe present-day geological processes, such as the horizontal motions at plate boundaries and vertical motions due to isostatic rebound.

Various models of plate motions have been constructed from different data sources. Rotation poles, calculated from transform faults, fault plane solutions, and by matching magnetic anomalies across spreading ridges, formed the basis of the NUVEL (DeMets et al., 1990) and MORVEL (DeMets et al., 2010) models (i.e., mid-ocean ridge velocities) of relative plate velocities at plate margins. Observations at continuously recording GPS stations and networks on all the continents were obtained in the recent velocity REVEL model (Sella et al., 2002). Exclusively GPS results were used in the GSRM (Kreemer et al., 2014) models of plate velocities and global strain rates in plate boundary zones. The GPS and sea-floor spreading models are compatible, but there are slight differences between them. GPS measurements give the present-day plate motion vectors, whereas ridge-analysis models deliver rotation poles averaged over a few million years, as represented by the magnetic and geological features at the ridges.

Exceptionally accurate measurements of changes in the distance between two stations on Earth are not restricted to GPS methods. For example, very long baseline interferometry (VLBI) observations made over several years in the early 1990s showed that the distances between pairs of stations on opposite sides of the Atlantic Ocean are increasing slowly at a mean rate of 17 mm yr^{-1} (Fig. 2.12). This figure is close to the long-term value of about 20 mm yr^{-1} interpreted from models of current plate motions (Fig. 2.7).

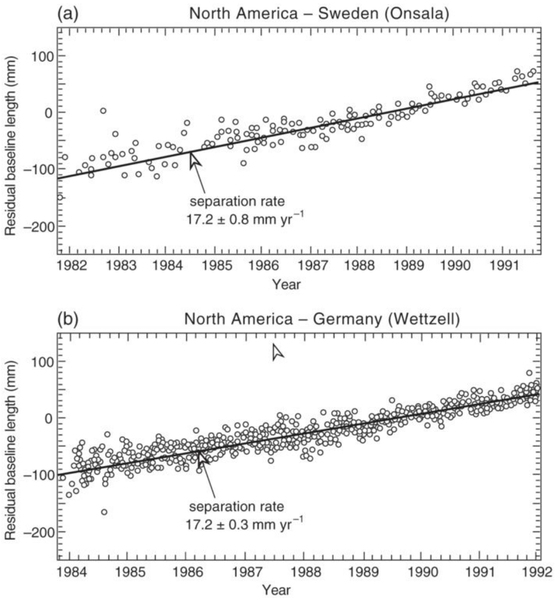


Fig. 2.12 Changes in separation between Westcott (Massachusetts, USA) and (a) Onsala (Sweden) and (b) Wettzell (Germany), as determined by very long baseline interferometry.

Source: J.W. Ryan, et al., Global scale tectonic plate motions with CDP VLBI data, in: *Contributions to Space Geodesy: Crustal Dynamics*, D.E. Smith and D.L. Turcotte, eds. Geodynamics Series, v.23, pp. 37–50, Washington, DC: AGU, 1993.

2.6 Plate Margin Features

It is important to keep in mind that the tectonic plates are not crustal units. They involve the entire thickness of the lithosphere, of which the crust is only the outer skin. Oceanic lithosphere is thin close to a ridge axis, but thickens with distance from the ridge, reaching a value of 80–100 km; the oceanic crust makes up only the top 5–10 km. Continental lithosphere may be up to 150 km thick, of which only the top 30–60 km is continental crust. Driven by dynamic processes within the Earth, the lithospheric plates move systematically relative to each other. This knowledge supplies the "missing link" in Wegener's continental drift hypothesis, removing one of the most serious objections to it. It is not necessary for the continents to plow through the rigid ocean basins; they are transported passively on top of the moving plates, as logs float on a stream. Continental drift is thus a consequence of plate motions.

The plate tectonic model involves the formation of new lithosphere at a ridge and its destruction at a subduction zone (Fig. 2.13). Since the mean density of oceanic lithosphere exceeds that of continental lithosphere, oceanic lithosphere can be subducted under continental or oceanic lithosphere, whereas continental lithosphere cannot underride oceanic lithosphere. Just as logs pile up where a stream dives under a surface obstacle, a continent that is transported into a subduction zone collides with the deep-sea trench, island arc, or adjacent continent. Such a collision results in an orogenic belt. In a continent–continent collision, neither plate can easily subduct, so relative plate motion may come to a halt. Alternatively, subduction may start at a new location behind one of the continents, leaving a mountain chain as evidence of the *suture zone* between the original colliding continents. The Alpine–Himalayan and Appalachian mountain chains are thought to have formed by this mechanism, the former in Tertiary times, the latter in several stages during the

Paleozoic. Plate tectonic theory is supported convincingly by an abundance of geophysical, petrological, and geological evidence from the three types of plate margin.

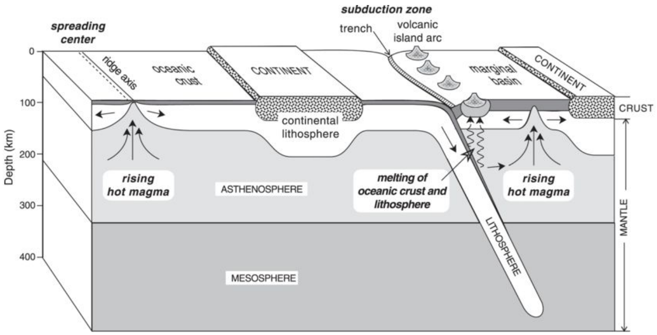

subduction zone
spreading center
trench
volcanic island arc
ridge axis
oceanic crust
CONTINENT
marginal basin
CONTINENT
CRUST
continental lithosphere
rising hot magma
ASTHENOSPHERE
melting of oceanic crust and lithosphere
rising hot magma
MANTLE
LITHOSPHERE
MESOSPHERE
Depth (km)
0
100
200
300
400

Fig. 2.13 Hypothetical vertical cross-section through a lithospheric plate from a spreading center to a subduction zone.

2.6.1 Constructive Margins

Although the ridges and rises are generally not centrally located in the ocean basins, they are often referred to as mid-ocean ridges. The type of oceanic basalt that is produced at an oceanic spreading center is even called a mid-ocean ridge basalt (MORB). Topographically, slow-spreading ridges have a distinct axial rift valley, which is missing on faster-spreading ridges. Partially molten upper mantle rocks (generally assumed to be peridotites) from the asthenosphere rise under the ridges. The decrease of pressure due to the changing depth causes further melting and the formation of basaltic magma. Their chemical compositions and the concentrations of long-lived radioactive isotopes suggest that MORB lavas are derived by fractionation (i.e., separation of components, perhaps by precipitation or crystallization) from the upwelling peridotitic mush. Differentiation is thought to take place at about the depth of the lower crustal gabbroic layer beneath the ridge in a small, narrow magma chamber. Some of the fluid magma extrudes near the central rift or ridge axis and flows as lava across the ocean floor; part is intruded as dikes and sills into the thin oceanic crust. The Vine–Matthews–Morley hypothesis for the origin of oceanic magnetic anomalies requires fairly sharp boundaries between alternately magnetized blocks of oceanic crust. This implies that the zone of dike injection is narrow and close to the ridge axis.

The distribution of earthquakes defines a narrow band of seismic activity close to the crest of an oceanic ridge. These earthquakes occur at shallow depths of a few kilometers and are mostly small; magnitudes of 6 or greater are rare. The seismic energy released at ridges is an insignificant part of the world-wide annual release. Analyses show that the earthquakes are associated with normal faulting, implying extension away from the ridge axis (see Section 7.2.5).

Heat flow in the oceans is highest at the ocean ridges and decreases systematically with distance away from the ridge. The thermal data conform to the model of sea-floor spreading. High axial values are caused by the formation of new lithosphere from the hot uprising magma at the ridge axis. The associated volcanism on the floor of the axial rift zones has been observed directly from deep-diving submersibles. With time, the lithosphere spreads away from the ridge and gradually cools, so that the heat outflow diminishes with increasing age or distance from the ridge.

Oceanic crust is thin, so the high-density mantle rocks occur at shallower depths than under the continents. This causes a general increase of the Earth's gravity field over the oceans, giving positive gravity anomalies. However, over the ridge systems gravity decreases toward the axis so that a "negative" anomaly is superposed on the normally positive oceanic gravity anomaly. The effect is due to the local density structure under the ridge. It has been interpreted in terms of anomalous mantle material with density slightly less than normal. The density is low because of the different mantle composition under the ridges and its high temperature.

The interpretation of magnetic anomalies formed by sea-floor spreading at constructive margins has already been discussed. The results provide direct estimates of the mean rates of plate motions over geological time intervals.

2.6.2 Destructive Margins

Subduction zones are found where a plate plunges beneath its neighbor to great depths, until pressure and temperature cause its consumption. This usually happens within a few hundred kilometers, but seismic tomography (Section 7.2.6) has shown that some descending slabs may sink to great depths, even to the core–mantle boundary. Density determines that the descending plate at a subduction zone is an oceanic one. The surface manifestation depends on the type of overriding plate. When this is another oceanic plate, the subduction zone is marked by a volcanic island arc and, parallel to it, a deep trench. The island arc lies near the edge of the overriding plate and is convex toward the underriding plate. The trench marks where the underriding plate turns down into the mantle (Fig. 2.13). It may be partly filled with carbonaceous and detrital sediments. Island arc and trench are a few hundred kilometers apart. Several examples are seen around the west and northwest margins of the Pacific plate (Fig. 2.7). Melting of the downgoing slab produces magma that rises to feed the volcanoes.

The intrusion of magma behind an island arc produces a back-arc basin on the inner, concave side of the arc. These basins are common in the Western Pacific. If the arc is close to a continent, the off-arc magmatism may create a marginal sea, such as the Sea of Japan. Back-arc basins and marginal seas are floored by oceanic crust.

A fine example of where the overriding plate is a continental one is seen along the west coast of South America. Compression between the Nazca and South American plates has generated the Andes, an arcuate-folded mountain belt near the edge of the continental plate. Active volcanoes along the mountain chain emit a type of lava, called andesite, which has a higher silica content than oceanic basalt. It does not originate from the asthenosphere type of magma. A current theory is that it may form by melting of the subducting slab and overriding plate at great depths. If some siliceous sediments from the deep-sea trench are carried down with the descending slab, they might enhance the silica content of the melt, producing a magma with andesite-type composition.

The seismicity at a subduction zone provides the key to the processes active there. Where one plate is thrust over the other, the shear causes hazardous earthquakes at shallow depths. Below this region, earthquakes are systematically distributed within the subducting plate. They form an inclined *Wadati–Benioff seismic zone*, which may extend for several hundred kilometers into the mantle. The deepest earthquakes have been registered down to about 700 km.

Studies of the focal mechanisms (Section 7.2.5) show that at shallow depths the downgoing plate is in a state of down-dip extension (Fig. 2.14a). Subducting lithosphere is colder and denser than the underlying asthenosphere. This gives it negative buoyancy, which causes it to sink, pulling the plate downward. At greater depths the mantle is more rigid than the asthenosphere, and its strength resists penetration (Fig. 2.14b). While the upper part is sinking, the bottom part is being partly supported by the deeper layers; this results in down-dip compression in the lower part of the descending slab and down-dip extension in the upper part. A gap in the depth distribution of seismicity may arise where the deviatoric stress changes from extensional to compressional. In a very deep subduction zone the increase of resistance with depth causes down-dip compression throughout the descending slab (Fig. 2.14c). In some cases part of the slab may break off and sink to great depths, where the earthquakes have compressional-type mechanisms (Fig. 2.14d); a gap in seismicity exists between the two parts of the slab.

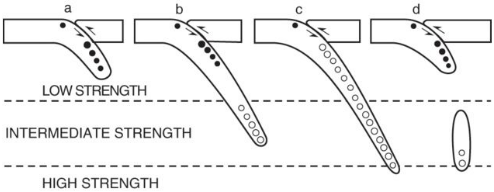


Fig. 2.14 Stresses acting on a subducting lithospheric plate. Arrows

indicate shear where the underriding plate is bent downward. Solid and open circles within the descending slab denote extension and compression, respectively; the size of the circle represents qualitatively the seismic activity. In (a), (b), and (d) extensional stress in the upper part of the plate is due to the slab being pulled into low-strength asthenosphere. In (b) resistance of the more rigid layer under the asthenosphere causes compression within the lower part of the slab; if the plate sinks far enough (c), the stress becomes compressional throughout; in some cases (d), the deep part of the lower slab may break off.

Source: B. Isacks and P. Molnar, Mantle earthquake mechanisms and the sinking of the lithosphere, *Nature* 223, 1121–1124, 1969.

Heat flow at a destructive plate margin reflects to some extent the spreading history of the plate. The plate reaches its maximum age, and so has cooled the furthest, by the time it reaches a subduction zone. The heat flow values over deep ocean basins are uniformly low, but the values measured in deep-sea trenches are the lowest found in the oceans. In contrast, volcanic arcs and back-arc basins often have anomalously high heat flow due to the injection of fresh magma.

Gravity anomalies across subduction zones have several distinctive features. Seaward of the trench the lithosphere flexes upward slightly before it begins its descent, causing a weak positive anomaly; the presence of water or low-density sediments in a deep-sea trench gives rise to a strong negative gravity anomaly; and over the descending slab a positive anomaly is observed, due in part to the mineralogical conversion of subducted oceanic crust to higher-density eclogite.

Subduction zones have no particular magnetic signature. Close to an active or passive continental margin the contrast between the magnetic properties of oceanic and continental crust produces a magnetic anomaly, but this is not a direct result of the plate tectonic processes. Over marginal basins magnetic anomalies are not lineated except in some rare cases. This is because the oceanic crust in the basin does not originate by sea-floor spreading at a ridge, but by diffuse intrusion throughout the basin.

2.6.3 Conservative Margins

Transform faults are strike–slip faults with steeply dipping fault planes. They may link segments of subduction zones, but they are mostly observed at constructive plate margins, where they connect oceanic ridge segments. Transform faults are the most seismically active parts of a ridge system, because here the relative motion between neighboring plates is most pronounced. Seismic studies have confirmed that the displacements on transform faults agree with the relative motion between the adjacent plates.

The trace of a transform fault may extend away from a ridge on both sides as a *fracture zone*. Fracture zones are among the most dramatic features of ocean-floor topography. Although only some tens of kilometers wide, a fracture zone can be thousands of kilometers long. It traces the arc of a small circle on the surface of the globe. This important characteristic allows fracture zones to be used for the deduction of relative plate motions, which cannot be obtained from the strike of a ridge or trench segment, where oblique spreading or subduction is possible (note, for example, the direction of plate convergence relative to the strike of the Aleutian island arc in Fig. 2.7).

Any displacement on the surface of a sphere is equivalent to a small rotation about a pole. The motion of one plate relative to the other takes place as a rotation about the *Euler pole* of relative rotation between the plates (see Section 2.9). This pole can be located from the orientations of fracture zones, because the strike of a transform fault is parallel to the relative motion between two adjacent plates. Thus a great circle normal to a transform fault or fracture zone must pass through the Euler pole of relative rotation between the two plates. If several great circles are drawn at different places on the fracture zone (or normal to different transform faults offsetting a ridge axis) they intersect at the Euler pole. Models of relative plate motions are obtained by determining the Euler poles of rotation between pairs of plates using magnetic anomalies, the directions of slip on earthquake fault planes at plate boundaries, and the topography that defines the strikes of transform faults. The rates of relative motion at different places on the plate boundaries (Fig. 2.7) were computed from the rates of rotation about the appropriate Euler poles.

There may be a large change in elevation across a fracture zone; this is related to the different thermal histories of the plates it separates. As a plate cools, it becomes more dense and less buoyant, so that it gradually sinks. Consequently, the depth to the top of the oceanic lithosphere increases with age, i.e., with distance from the spreading center. Places facing each other across a transform fault are at different distances from their respective spreading centers. They have different ages and so have subsided by different amounts relative to the ridge. This may result in a noticeable elevation difference across the fracture zone.

Ultrabasic rocks are found in fracture zones and there may be local magnetic anomalies. Otherwise, the magnetic effect of a transform fault is to interrupt the oceanic magnetic lineations parallel to a ridge axis, and to offset them in the same measure as ridge segments. This results in a very complex pattern of magnetic lineations in some ocean basins (e.g., in the northeast Pacific).

A transform fault can also connect subduction zones. Suppose a consuming plate boundary consisted originally of two opposed subduction zones (Fig. 2.15a). Plate Y is consumed below plate X along the segment *ab* of the boundary, whereas plate X is consumed beneath plate Y along segment *bc*. The configuration is unstable, because a trench cannot sustain subduction in opposite directions. Consequently, a dextral transform fault develops at point *b*. After some time, motion on the fault displaces the lower segment to the position *b′c* (Fig. 2.15b). An example of such a transform boundary is the Alpine fault in New Zealand (Fig. 2.15c). To the northeast of the North Island, the Pacific plate is being subducted at the Tonga–Kermadec trench. To the southwest of the South Island, the Pacific plate overrides the Tasman Sea at the anomalous Macquarie Ridge (earthquake analysis has shown that the plate margin at this ridge is compressive; the compression may be too slow to allow a trench to develop). The Alpine fault linking the two opposed subduction zones is therefore a dextral transform fault.

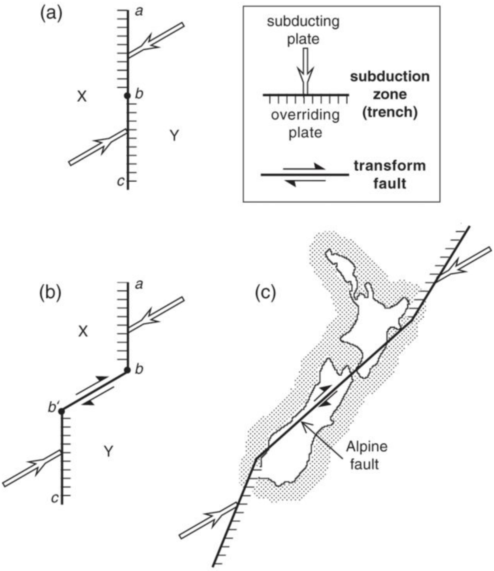

(a)
a
X
b
Y
c
subducting plate
subduction zone (trench)
overriding plate
transform fault
(b)
a
X
b
b′
Y
c
(c)
Alpine fault

Fig. 2.15 (a) A consuming plate boundary consisting of two opposed subduction zones; along *ab* plate Y is consumed below plate X and along *bc* plate X is consumed beneath plate Y. (b) Development of a transform fault which displaces *bc* to the position *b′c*. (c) The Alpine fault in New Zealand is an example of such a transform boundary.

Source: D.P. McKenzie and W.J. Morgan, Evolution of triple junctions, *Nature* 224, 125–133, 1969.

2.6.4 Diffuse Margins

The most striking feature of the global seismicity map (Fig. 2.6) is the concentration of earthquakes in well-defined narrow zones that define plate margins. There are, however, large areas, hundreds of kilometers across, in which earthquakes occur frequently but which are broadly distributed. This is evident in Fig. 7.5, which shows more than one million earthquakes. For example, a roughly equatorial, east–west pattern of earthquakes between India and Australia marks a diffuse margin between the Indian and Australian plates, which have different rotation poles. The rates of relative plate motions at diffuse margins are a fraction of those at other plate margins. In contrast to them, a diffuse margin appears to be a broad zone of deformation resulting from the differential rotation between the adjacent plates.

The geophysical deployment of GPS networks has made it possible to measure in only a few weeks or months the relative motions that are currently taking place at plate margins. The results complement other geophysical evidence and, moreover, indicate the locations of possible additional plate margins (dashed lines in Fig. 2.7), for example in eastern Asia and Africa. The East African Rift Zone, for example, is actively volcanic and opening at the rate of a few millimeters per year. This is slower than the rates at most active margins, but the spreading indicates the probable location of a constructive plate boundary that divides the African plate in two, to form a western Nubia plate and an eastern Somalia plate. The Somalia plate may be bordered on its southern margin by a postulated Lwandle plate, although the diffuse boundaries of this plate are not well delineated. More than 50 tectonic plates have been named, of which seven are major plates and a further dozen or so are well established using geophysical and geological criteria.

2.7 Triple Junctions

It is common, although imprecise, to refer to a plate margin by its dominant topographic feature, rather than by the nature of the margin. A ridge (R) represents a constructive margin or spreading center, a trench (T) refers to a destructive margin or subduction zone, and a transform fault (F) stands for a conservative margin. Each margin is a location where two tectonic plates adjoin. Inspection of the plate margins in Fig. 2.7 shows that there are several places where three plates come together, but none where four or more plates meet. The meeting points of three plate boundaries are called *triple junctions*. They are important in plate tectonics because the relative motions between the plates that form a triple junction are not independent. This may be appreciated by considering the plate motions in a small plane surrounding the junction.

Consider the plate velocities at an RTF junction formed by all three types of boundary (Fig. 2.16a). If the plates are rigid, their relative motions take place entirely at their margins. Let ${}_{\mathrm{A}}\boldsymbol{V}_{\mathrm{B}}$ denote the velocity of plate B relative to plate A, ${}_{\mathrm{B}}\boldsymbol{V}_{\mathrm{C}}$ the velocity of plate C relative to plate B, and ${}_{\mathrm{C}}\boldsymbol{V}_{\mathrm{A}}$ the velocity of plate A relative to plate C. Note that these quantities are vectors; their directions are as important as their magnitudes. They can be represented on a vector diagram by straight lines with directions parallel to and lengths proportional to the velocities. In a circuit about the triple junction, an observer must return to the starting point. Thus, a vector diagram of the interplate velocities is a closed triangle (Fig. 2.16b). The velocities are related by

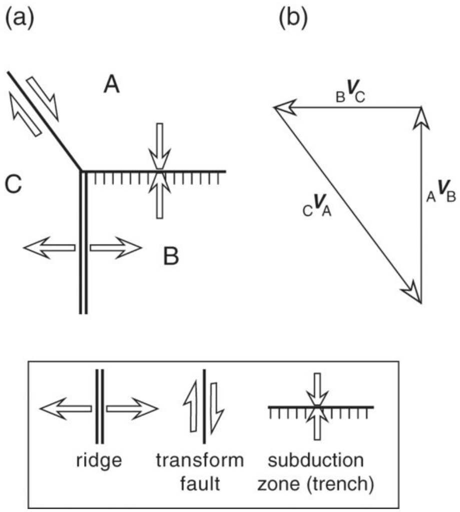

(a)
(b)
A
C
B
$_B\boldsymbol{V}_C$
$_C\boldsymbol{V}_A$
$_A\boldsymbol{V}_B$
ridge
transform fault
subduction zone (trench)

Fig. 2.16 (a) Triple junction formed by a ridge, trench, and transform fault; and (b) vector diagram of the relative velocities at the three boundaries.

Source: D.P. McKenzie and R.L. Parker, The North Pacific: an example of tectonics on a sphere, *Nature* 216, 1276–1280, 1967.

$$ {}_{\mathrm{A}}V_{\mathrm{B}} + {}_{\mathrm{B}}V_{\mathrm{C}} + {}_{\mathrm{C}}V_{\mathrm{A}} = 0. \quad (2.1) $$

This planar model is a “flat Earth” representation. As discussed in Section 2.9, displacements on the surface of a sphere are rotations about Euler poles of relative motion. This can be taken into account by replacing each linear velocity $\boldsymbol{V}$ in Eq. (2.1) by the rotational velocity $\boldsymbol{\omega}$ about the appropriate Euler pole.

2.7.1 Stability of Triple Junctions

The different combinations of three plate margins define 10 possible types of triple junction. The combinations correspond to all three margins being of one type (RRR, TTT, FFF), two of the same type and one of the other (RRT, RRF, FFT, FFR, TTR, TTF), and all different (RTF). Different combinations of the sense of subduction at a trench increase the number of possible junctions to 16. Not all of these junctions are stable in time. For a junction to preserve its geometry, the orientations of the three plate boundaries must fulfill conditions that allow the relative velocities to satisfy Eq. (2.1). If they do so, the junction is stable and can maintain its shape. Otherwise, the junction is unstable and must evolve in time to a stable configuration.

The stability of a triple junction is assessed by considering how it can move along any of the plate boundaries that form it. The velocity of a plate can be represented by its coordinates in *velocity space*. Consider, for example, a trench or consuming plate margin (Fig. 2.17a). Point A in velocity space represents the consuming plate, which has a larger velocity than B for the overriding plate. A triple junction in which one plate margin is a trench can lie anywhere on this boundary, so the locus of its possible velocities is a line *ab* parallel to the trench. The trench is fixed relative to the overriding plate B, so the line *ab* must pass through B. Similar reasoning shows that a triple junction on a transform fault is represented in velocity space by a line *ab* parallel to the fault and passing through both A and B (Fig. 2.17b). A triple junction on a ridge gives a velocity line *ab* parallel to the ridge; in the case of symmetrical spreading normal to the trend of the ridge the line *ab* is the perpendicular bisector of AB (Fig. 2.17c).

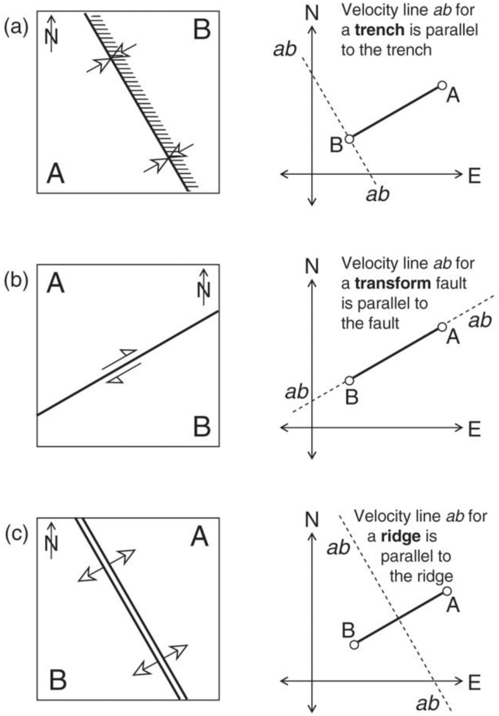

(a)
N
B
A
N
E
ab
ab
Velocity line *ab* for a **trench** is parallel to the trench
A
B
(b)
A
N
B
N
E
ab
ab
Velocity line *ab* for a **transform** fault is parallel to the fault
A
B
(c)
N
A
B
N
E
ab
ab
Velocity line *ab* for a **ridge** is parallel to the ridge
A
B

Fig. 2.17 Plate margin geometry (*left*) and locus *ab* of a triple junction in velocity space (*right*) for (a) a trench, (b) a transform fault, and (c) a ridge.

Source: A. Cox and R.B. Hart, *Plate Tectonics*, 392 pp., Boston: Blackwell Scientific Publications, 1986.

Now consider the RRR type of triple junction, formed by three ridges (Fig. 2.18a). The locus of the triple junction on the ridge between any pair of plates is the perpendicular bisector of the corresponding side of the velocity triangle ABC. The perpendicular bisectors of the sides of a triangle always meet at a point (the circumcenter). In velocity space this point satisfies the velocities on all three ridges simultaneously, so the RRR triple junction is always stable. Conversely, a triple junction formed by three intersecting transform faults (FFF) is always unstable, because the velocity lines form the sides of a triangle, which can never meet in a point (Fig. 2.18b). The other types of triple junction are conditionally stable, depending on the angles between the different margins. For example, in an RTF triple junction the velocity lines of the trench *ac* and transform fault *bc* must both pass through C, because this plate is common to both boundaries. The junction is stable if the velocity line *ab* of the ridge also passes through C, or if the trench and transform fault have the same trend (Fig. 2.18c). By similar reasoning, the FFT triple junction is only stable if the trench has the same trend as one of the transform faults (Fig. 2.18d).

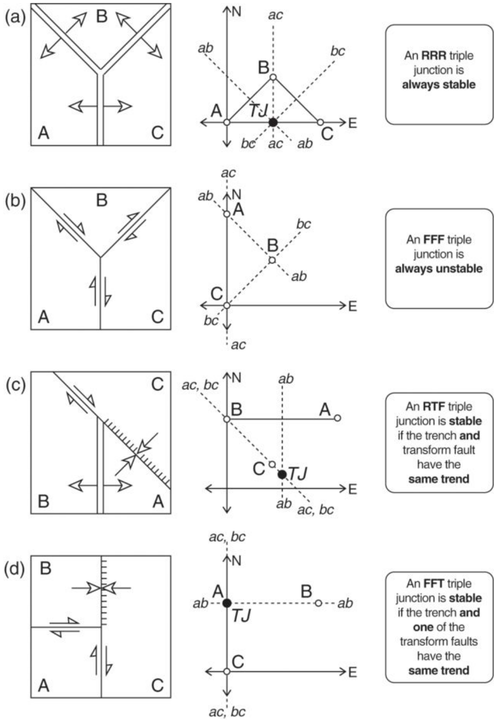

(a)
B
A
C
N
E
ac
ab
bc
TJ
An **RRR** triple junction is **always stable**
(b)
An **FFF** triple junction is **always unstable**
(c)
ac, bc
An **RTF** triple junction is **stable** if the trench **and** transform fault have the **same trend**
(d)
An **FFT** triple junction is **stable** if the trench **and one** of the transform faults have the **same trend**

Fig. 2.18 Triple junction configuration (*left*), velocity lines of each margin in velocity space (*center*), and stability criteria (*right*) for selected triple junctions *TJ* .

Source: A. Cox and R.B. Hart, *Plate Tectonics*, 392 pp., Boston: Blackwell Scientific Publications, 1986.

In the present phase of plate tectonics only a few of the possible types of triple junction appear to be active. An RRR type is formed where the Galapagos Ridge meets the East Pacific Rise at the junction of the Cocos, Nazca, and Pacific plates. The only known TTT type of junction is formed by the intersection of the Japan and Izu-Bonin trenches with the Nankai-Sagami Trough. The San Andreas fault in California terminates in an FFT type of junction at its northern end, where it joins the Mendocino Fracture Zone.

2.7.2 Evolution of Triple Junctions in the Northeast Pacific

Oceanic magnetic anomalies in the northeast Pacific form a complex striped pattern. The anomalies can be identified by interpreting their shapes. Their ages can be found by comparison with a geomagnetic polarity timescale such as that shown in Fig. 12.32, which gives the age of each numbered chron since the Late Jurassic. In the northeast Pacific the anomalies become younger toward the North American continent in the east, and toward the Aleutian trench in the north. The anomaly pattern produced at a ridge is usually symmetric (as in Fig. 2.9), but in the northeast Pacific only the western half of an anomaly pattern is observed. The plate on which the eastern half of the anomaly pattern was formed is called the Farallon plate. It and the ridge itself are largely missing and have evidently been subducted under the American plate. Only two small remnants of the Farallon plate still exist: the Juan de Fuca plate off the coast of British Columbia, and the Rivera plate at the mouth of the Gulf of California. The magnetic anomalies also indicate that another plate, the Kula plate, existed in the Late Mesozoic but has now been entirely consumed under Alaska and the Aleutian trench. The anomaly pattern shows that in the Late Cretaceous the Pacific, Kula, and Farallon plates were diverging from each other and thus met at an RRR type triple junction. This type of junction is stable and preserved its shape during subsequent evolution of the plates. It is therefore possible to reconstruct the relative motions of the Pacific, Kula, and Farallon plates in the Cenozoic (Fig. 2.19a–c).

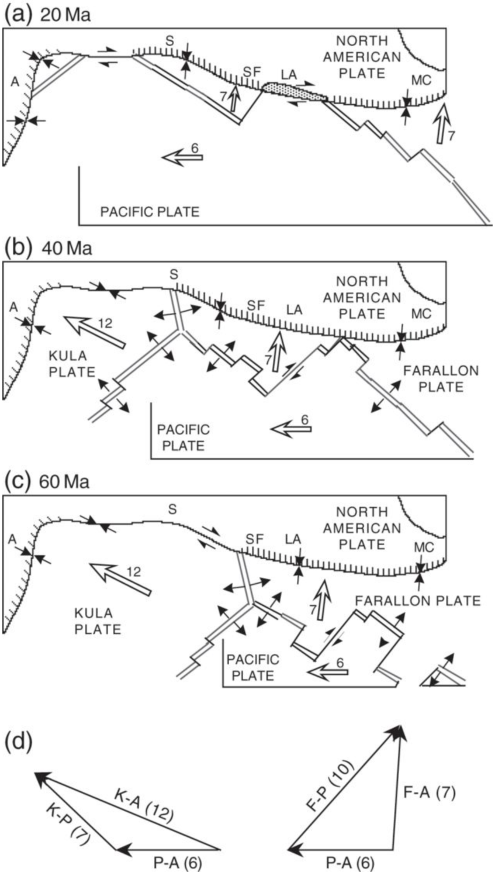

(a) 20 Ma
S
SF
LA
NORTH
AMERICAN
PLATE
MC
A
7
6
7
PACIFIC PLATE
(b) 40 Ma
S
SF
LA
NORTH
AMERICAN
PLATE
MC
A
12
KULA
PLATE
7
FARALLON
PLATE
PACIFIC
PLATE
6
(c) 60 Ma
S
SF
LA
NORTH
AMERICAN
PLATE
MC
A
12
KULA
PLATE
7
FARALLON PLATE
PACIFIC
PLATE
6
(d)
K-A (12)
K-P (7)
P-A (6)
F-P (10)
F-A (7)
P-A (6)

Fig. 2.19 (a)–(c) Extrapolated plate relationships in the northeast Pacific at different times in the Cenozoic. Letters on the American plate give approximate locations of some modern cities for reference: MC, Mexico City; LA, Los Angeles; SF, San Francisco; S, Seattle; A, Anchorage. The shaded area in (a) is an unacceptable overlap. (d) Vector diagrams of the relative plate velocities at the Kula–Pacific–American and Farallon–Pacific–American triple junctions (numbers are velocities in cm yr^{-1} relative to the American plate).

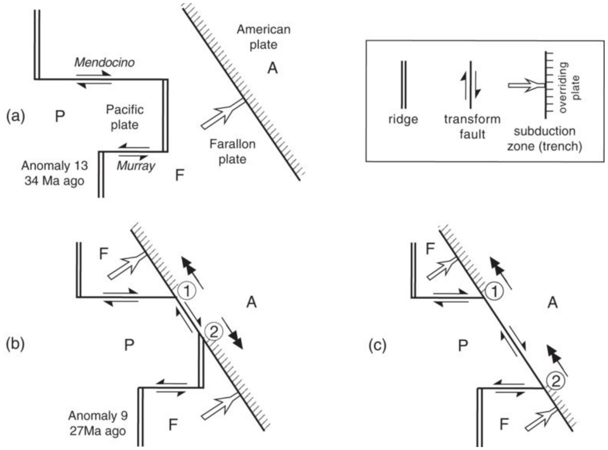
Source: T. Atwater, Implications of plate tectonics for the Cenozoic tectonic evolution of western North America, *Geol. Soc. Amer. Bull.* 81, 3513–3536, 1970.

The anomaly ages are known from the magnetic timescale, so the anomaly spacing allows the half-rates of spreading to be determined.

In conjunction with the trends of fracture zones, the anomaly data give the rates and directions of spreading at each ridge. The anomaly pattern at the mouth of the Gulf of California covers the last 4 Myr and gives a mean half-rate of spreading of 3 cm yr^{-1} parallel to the San Andreas fault. This indicates that the Pacific plate has moved northward past the American plate at this boundary with a mean relative velocity of about 6 cm yr^{-1} during the last 4 Ma. The half-rate of spreading on the remnant of the Farallon–Pacific ridge is 5 cm yr^{-1}, giving a relative velocity of 10 cm yr^{-1} between the plates. A vector diagram of relative velocities at the Farallon–Pacific–American triple junction (Fig. 2.19d) shows convergence of the Farallon plate on the American plate at a rate of 7 cm yr^{-1}. Similarly, the spacing of east–west trending magnetic anomalies in the Gulf of Alaska gives the half-rate of spreading on the Kula–Pacific ridge, from which it may be inferred that the relative velocity between the plates was 7 cm yr^{-1}. A vector diagram combining this value with the 6 cm yr^{-1} northward motion of the Pacific plate gives a velocity of 12 cm yr^{-1} for the Kula plate relative to the American plate.

Using these velocities, the history of plate evolution in the Cenozoic can be deduced by extrapolation. The interpretation is tenuous, as it involves unverifiable assumptions. The most obvious is that the Kula–Pacific motion in the Late Cretaceous (80 Myr ago) and the American–Pacific motion of the past 4 Myr have remained constant throughout the Cenozoic. With this proviso, it is evident that triple junctions formed and migrated along the American plate margin. The Kula–American–Farallon RTF junction was slightly north of the present location of San Francisco 60 Myr ago (Fig. 2.19 c); it moved to a position north of Seattle 20 Myr ago (Fig. 2.19a). Around that time in the Oligocene an FFT junction formed between SF and LA, while the Farallon–Pacific–American RTF junction evolved to the south. The development of these two triple junctions is due to the collision and subduction of the Farallon–Pacific ridge at the Farallon–American trench.

At the time of magnetic anomaly 13, about 34 Myr ago, a north–south striking ridge joined the Mendocino and Murray transform faults as part of the Farallon–Pacific plate margin to the west of the American trench (Fig. 2.20a). By the time of anomaly 9, about 27 Myr ago, the ridge had collided with the trench and been partly consumed by it (Fig. 2.20b). The Farallon plate now consisted of two fragments: an FFT junction developed at point 1, formed by the San Andreas fault system, the Mendocino fault, and the consuming trench to the north; and an RTF junction formed at point 2. Both junctions are stable when the trenches are parallel to the transform fault along the San Andreas system. Analysis of the velocity diagrams at each triple junction shows that point 1 migrated to the northwest and point 2 migrated to the southeast at this stage. Later, when the southern segment of the Farallon–Pacific ridge had been subducted under the American plate, the Murray transform fault changed the junction at point 2 to an FFT junction, which has subsequently also migrated to the northwest (Fig. 2.20c).



(a)
P
Mendocino
Pacific plate
Anomaly 13
34 Ma ago
Murray
F
American plate
A
Farallon plate
ridge
transform fault
subduction zone (trench)
overriding plate
(b)
F
①
②
A
P
Anomaly 9
27Ma ago
F
(c)
F
①
A
P
②
F

Fig. 2.20 Formation of the San Andreas fault as a result of the evolution of triple junctions in the northeast Pacific during the Oligocene: plate geometries at the times of (a) magnetic anomaly 13, about 34 Myr ago, (b) anomaly 9, about 27 Myr ago, and (c) further development when the Murray fracture zone collides with the trench. Double-headed arrows show directions of migration of

triple junctions 1 and 2 along the consuming plate margin.

Source: D.P. McKenzie and W.J. Morgan, Evolution of triple junctions, *Nature* 224, 125–133, 1969.

2.8 Hotspots

In 1958 S. W. Carey coined the term “hot spot” – now often reduced to “hotspot” – to refer to a long-lasting center of surface volcanism and locally high heat flow. At one time more than 120 of these thermal anomalies were proposed. Application of more stringent criteria has reduced their number to about 41 (Fig. 2.21), although some of these may fail more stringent criteria. The hotspots can occur on the continents (e.g., Yellowstone), but are more common in the ocean basins. The oceanic hotspots are associated with depth anomalies. If the observed depth is compared with the depth predicted by cooling models of the oceanic lithosphere, the hotspots are found to lie predominantly in broad, shallow regions, where the lithosphere apparently swells upward. This elevates denser mantle material, which creates a mass anomaly and disturbs the geoid; the effect is partially mitigated by reduced density of material in the hot, rising plume. The geoid surface is also displaced by subduction zones. The residual geoid obtained by removing the effects associated with cold subducting slabs shows a remarkable correlation with the distribution of hotspots (Fig. 2.21). The oceanic hotspots are found in conjunction with intraplate island chains, which provide clues to the origin of hotspots and allow them to be used for measuring geodynamic processes.

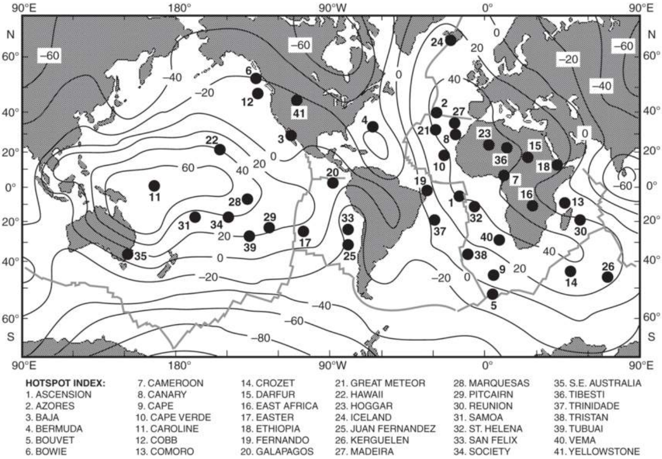

90°E
180°
90°W
0°
90°E
N
60°
40°
20°
0°
20°
40°
60°
S
HOTSPOT INDEX:
1. ASCENSION
2. AZORES
3. BAJA
4. BERMUDA
5. BOUVET
6. BOWIE
7. CAMEROON
8. CANARY
9. CAPE
10. CAPE VERDE
11. CAROLINE
12. COBB
13. COMORO
14. CROZET
15. DARFUR
16. EAST AFRICA
17. EASTER
18. ETHIOPIA
19. FERNANDO
20. GALAPAGOS
21. GREAT METEOR
22. HAWAII
23. HOGGAR
24. ICELAND
25. JUAN FERNANDEZ
26. KERGUELEN
27. MADEIRA
28. MARQUESAS
29. PITCAIRN
30. REUNION
31. SAMOA
32. ST. HELENA
33. SAN FELIX
34. SOCIETY
35. S.E. AUSTRALIA
36. TIBESTI
37. TRINIDADE
38. TRISTAN
39. TUBUAI
40. VEMA
41. YELLOWSTONE

Fig. 2.21 The global distribution of 41 hotspots and their relationship to the residual geoid obtained by correcting geoid heights (shown in meters above the reference ellipsoid) for the effects of cold subducting slabs.

Source: S.T. Crough and D.M. Jurdy, Subducted lithosphere, hotspots, and the geoid, *Earth Planet. Sci. Lett.* 48, 15–22, 1980.

Two types of volcanic island chains are important in plate tectonics. The chains of islands associated with deep oceanic trenches at consuming plate margins are related to the process of subduction and have an arcuate shape. Nearly linear chains of volcanic islands are observed within oceanic basins far from active plate margins. These intraplate features are particularly evident on a bathymetric map of the Pacific Ocean. The Hawaiian, Marquesas, Society, and Austral Islands form subparallel chains that trend approximately perpendicular to the axis of ocean-floor spreading on the East Pacific Rise. The most closely studied is the Hawaiian Ridge (Fig. 2.22a). The volcanism along this chain decreases from present-day activity at the southeast, on the island of Hawaii, to long-extinct seamounts and guyots toward the northwest along the Emperor Seamount chain. The history of development of the chain is typical of other linear volcanic island chains in the Pacific basin (Fig. 2.22b). It was explained in 1963 by J. T. Wilson, before the modern theory of plate tectonics was formulated.

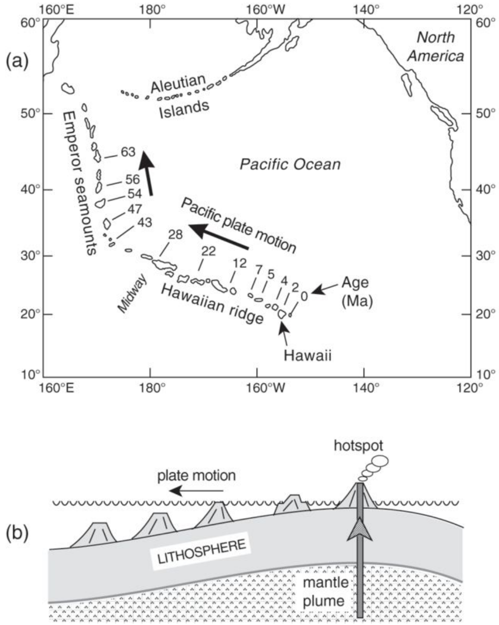

(a)
160°E
180°
160°W
140°
120°
60°
50°
40°
30°
20°
10°
North America
Aleutian Islands
Emperor seamounts
Pacific Ocean
Pacific plate motion
63
56
54
47
43
28
22
12
7
5
4
2
0
Age (Ma)
Midway
Hawaiian ridge
Hawaii
(b)
hotspot
plate motion
LITHOSPHERE
mantle plume

Fig. 2.22 (a) The Hawaiian Ridge and Emperor Seamount volcanic chains trace the motion of the Pacific plate over the Hawaiian hotspot; numbers give the approximate age of volcanism; note the change of direction about 43 Myr ago. (b) Sketch illustrating the formation of volcanic islands and seamounts as a lithospheric plate moves over a hotspot.

Sources: T.H. Van Andel, Seafloor spreading and plate tectonics, in: *Understanding the Earth*; C.J. Brown, C.J. Hawkesworth and R.C.L. Wilson, eds., pp. 167–186, Cambridge: Cambridge University Press, 1992.

A hotspot is a long-lasting magmatic center rooted in the mantle below the lithosphere. A volcanic complex is built up above the magmatic source, forming a volcanic island or, where the structure does not reach sea level, a seamount. The motion of the plate transports the island away from the hotspot and the volcanism becomes extinct. The upwelling material at the hotspot elevates the ocean floor by up to 1500 m above the normal depth of the ocean floor, creating a depth anomaly. As they move away from the hotspot, the by now extinct volcanic islands sink beneath the surface; some are truncated by erosion to sea level and become guyots. Coral atolls may accumulate on some guyots. The volcanic chain is aligned with the motion of the plate.

Confirmation of this theory is obtained from radiometric dating of basalt samples from islands and seamounts along the Hawaiian Ridge portion of the Hawaiian–Emperor chain. The basalts increase in age with distance from the active volcano Kilauea on the island of Hawaii (Fig. 2.23). The trend shows that the average rate of motion of the Pacific plate over the Hawaiian hotspot has been about 10 cm yr^{-1} during the last 20–40 Myr. The change in trend between the Hawaiian Ridge and the Emperor Seamount chain indicates a change in direction and speed of the Pacific plate about 43 Myr ago, at which time there was a global reorganization of plate motions. The earlier rate of motion along the Emperor chain is less well determined but is estimated to be about 6 cm yr^{-1}.

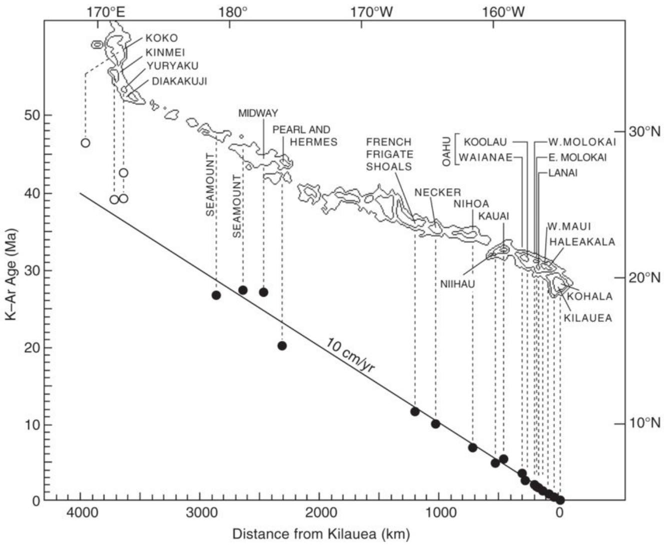


Fig. 2.23 Age of basaltic volcanism along the Hawaiian Islands as a function of distance from the active volcano Kilauea.

Source: G.B. Dalrymple, D.A. Clague and M.A. Lanphere, Revised age for Midway volcano, Hawaiian volcanic chain, *Earth Planet. Sci. Lett.* 37, 107–116, 1977.

Radiometric dating of linear volcanic chains in the Pacific basin gives almost identical rates of motion over their respective hotspots. This suggests that the hotspots form a stationary network, at least relative to the lithosphere. The velocities of plate motions over the hotspots are therefore regarded as absolute velocities, in contrast to the velocities derived at plate margins, which are the relative velocities between neighboring plates. The assumption that the hotspots are indeed stationary has been contested by studies that have yielded rates of inter-hotspot motion of the order of 1.5–2 cm yr^{-1} (comparable to present spreading rates in the Atlantic). Thus, the notion of a stationary hotspot reference frame may only be valid for a limited time interval. Nevertheless, any motions between hotspots are certainly much slower than the motions of plates, so the hotspot reference frame provides a useful guide to absolute plate motions over the typical time interval (~10 Myr) in which incremental sea-floor spreading is constant.

As well as geophysical evidence, there are geochemical anomalies associated with hotspot volcanism. The type of basalt extruded at a hotspot is different from the andesitic basalts formed in subduction zone magmatism. It also has a different petrology from the MORBs formed during sea-floor spreading and characteristic of the ocean floor. The hotspot source is assumed to be a mantle plume that reaches the surface. Mantle plumes are fundamental features of mantle dynamics, but they remain poorly understood. Although they are interpreted as long-term features, it is not known for how long they persist, or how they interact with convective processes in the mantle. Their role in heat transport and mantle convection, with consequent influence on plate motions, is believed to be important but is uncertain. Their sources are controversial. Some interpretations favor a comparatively shallow origin above the 660 km discontinuity, but the prevailing opinion appears to be that the plumes originate in the D″ layer at the core–mantle boundary. This requires the mantle plume to rise through the entire thickness of the mantle (see Fig. 9.28).

Seismic studies of the mantle have not yet been able to verify the existence of such features. However, the stationary nature of the hotspot network relative to the lithosphere provides a reference frame for determining absolute plate motions, and for testing the hypothesis of true polar wander, which is the long-term motion of the rotation axis relative to the mantle.

2.9 Plate Motion on the Surface of a Sphere

One of the great mathematicians of the eighteenth century was Leonhard Euler (1707–1783) of Switzerland. He made numerous fundamental contributions to pure mathematics, including complex numbers (Box 4.6) and spherical trigonometry (Box 2.1). A corollary of one of his theorems shows that the displacement of a rigid body on the surface of a sphere is equivalent to a rotation about an axis that passes through its center. This is applicable to the motion of a lithospheric plate.

Box 2.1 Spherical Trigonometry

The sides of a triangle on a plane surface are straight lines and the sum of its internal angles is 180° (or π radians). Let the angles be A, B, and C and the lengths of the sides opposite each of these angles be a, b, and c, as in Fig. B2.1.1a. The sizes of the angles and the lengths of the sides are governed by the *sine law*:

$$\frac{\sin A}{a} = \frac{\sin B}{b} = \frac{\sin C}{c}. \tag{1}$$

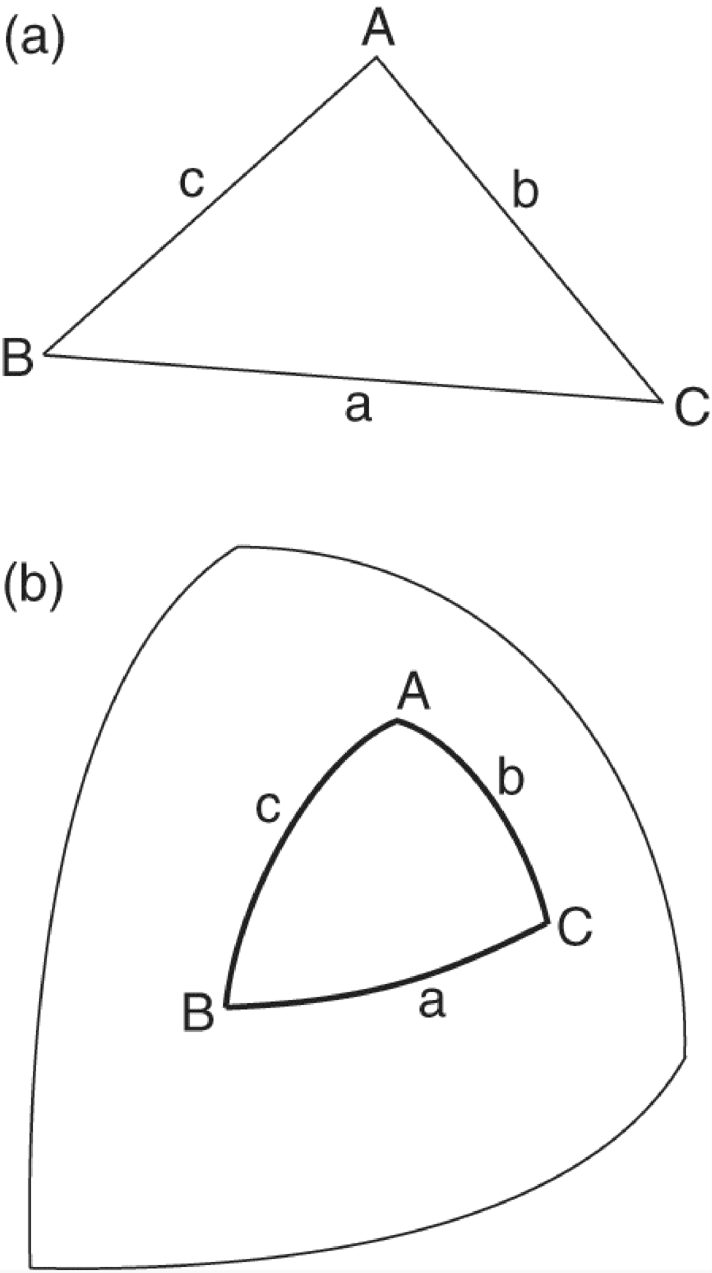


Fig. B2.1.1 The sides and angles of (a) a plane triangle, (b) a spherical triangle.

The length of any side is related to the lengths of the other two sides and to the angle they include by the *cosine law*, which for the side *a* is

$$a^2 = b^2 + c^2 - 2bc \cos A, \tag{2}$$

with similar expressions for the sides *b* and *c*.

The sides of a triangle on a spherical surface are great circle arcs and the sum of the internal angles is greater than 180°. The angle between two great circles at their point of intersection is defined by the tangents to the great circles at that point. Let the angles of a spherical triangle be *A, B*, and *C*, and let the lengths of the sides opposite each of these angles be *a, b* and *c*, respectively, as in Fig. B2.1.1b. The lengths of the sides may be converted to the angles they subtend at the center of the Earth. For example, the distance from pole to equator on the Earth's surface may be considered as 10,007 km or as 90 degrees of arc. Expressing the sides of the spherical triangle as angles of arc, the law of sines is

$$\frac{\sin A}{\sin a} = \frac{\sin B}{\sin b} = \frac{\sin C}{\sin c} \tag{3}$$

and the law of cosines is

$$\cos a = \cos b \cos c + \sin b \sin c \cos A. \tag{4}$$

Any motion restricted to the surface of a sphere takes place along a curved arc that is a segment of either a great circle (centered, like a "circle of longitude," at the Earth's center) or a small circle. Small circles are defined relative to a pole of rotational symmetry (such as the geographical pole, when we define "circles of latitude"). A point on the surface of the sphere can be regarded as the end-point of a radius vector from the center of the Earth to the point. Any position on the spherical surface can be specified by two angles, akin to latitude and longitude, or, alternatively, by direction cosines (Box 2.2). As a result of Euler's theorem, any displacement of a point along a small circle is equivalent to rotating the radius vector about the pole of symmetry, which is called the Euler pole of the rotation. A displacement along a great circle – the shortest distance between two points on the surface of the sphere – is a rotation about an Euler pole 90° away from the arcuate path. Euler poles were described in the discussion of conservative plate margins (Section 2.6.3); they play an important role in paleogeographic reconstructions using apparent polar wander paths (see Section 12.2.4.3).

Box 2.2 Direction Cosines

It is often useful to express a direction with the aid of *direction cosines*. These are the cosines of the angles that the direction makes with the reference axes. Define the z-axis along the Earth's spin axis, the x-axis along the Greenwich meridian, and the y-axis normal to both of these, as in Fig. B2.2.1. If a line makes angles α_x, α_y, and α_z to the x-, y-, and z-axes, respectively, its direction cosines with respect to these axes are

$$l = \cos\alpha_x; \quad m = \cos\alpha_y; \quad n = \cos\alpha_z. \tag{1}$$

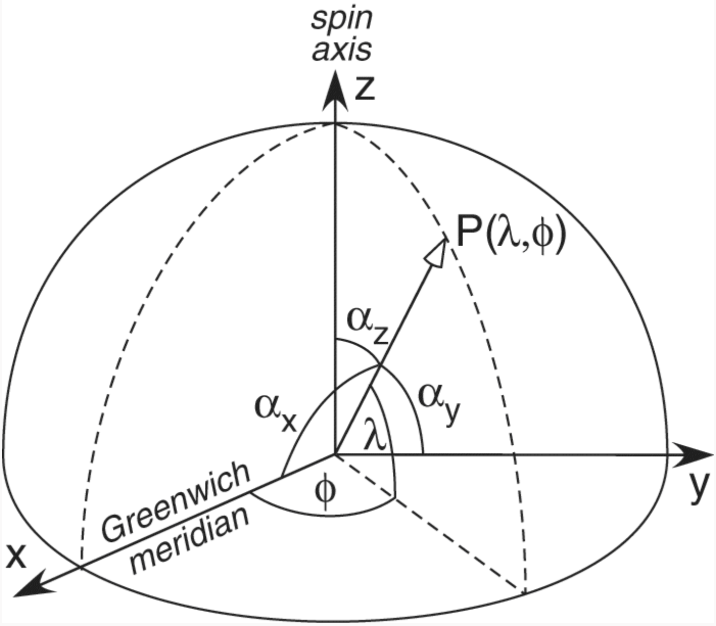


Fig. B2.2.1 The definition of direction cosines.

Consider a position P on the Earth's surface with latitude λ and longitude ϕ. A line of length R from the center of the Earth to the point P has projections $R\cos\alpha_z$ ($= R\sin\lambda$) on the z-axis and $R\sin\alpha_z$ ($= R\cos\lambda$) in the equatorial plane. The latter has projections ($R\cos\lambda\cos\phi$) and ($R\cos\lambda\ \sin\phi$) on the x- and y-axes, respectively. The direction cosines of the line are thus

$$\begin{aligned} l &= \cos\lambda\ \cos\phi \\ m &= \cos\lambda\ \sin\phi \\ n &= \sin\lambda. \end{aligned} \tag{2}$$

The angle Δ between two lines with direction cosines (l_1, m_1, n_1) and (l_2, m_2, n_2) is given by

$$\cos\Delta = l_1 l_2 + m_1 m_2 + n_1 n_2. \tag{3}$$

These relationships are useful for computing great circle distances and the angular relationships between lines.

2.9.1 Euler Poles of Rotation

Geophysical evidence does not in itself yield absolute plate motions. Present-day seismicity reflects relative motion between contiguous plates, oceanic magnetic anomaly patterns reveal long-term motion between neighboring plates, and paleomagnetism does not resolve displacements in longitude about a paleopole. The relative motion between plates is described by keeping one plate fixed and moving the other one relative to it; that is, we rotate it away from (or toward) the fixed plate (Fig. 2.24). The geometry of a rigid plate on the surface of a sphere is outlined by a set of bounding points, which maintain fixed positions relative to each other. Provided it remains rigid, each point of a moving plate describes an arc of a different small circle about the same Euler pole. Thus, the motion between plates is equivalent to a relative rotation about their mutual Euler rotation pole.

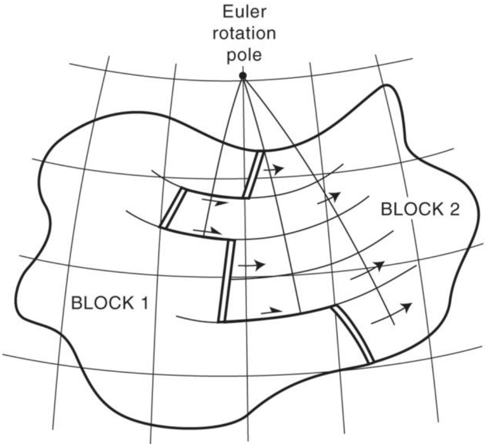


Fig. 2.24 The displacement of a rigid plate on the surface of a sphere is equivalent to the rotation of the plate about an Euler pole.

Source: W.J. Morgan, Rises, trenches, great faults, and crustal blocks, *J. Geophys. Res.* 73, 1959–1982, 1968.

The traces of past and present-day plate motions are recorded in the geometries of transform faults and fracture zones, which mark, respectively, the present-day and earlier locations of conservative plate margins. A segment of a transform fault represents the local path of relative motion between two plates. As such, it defines a small circle about the Euler pole of relative rotation between the plates. Great circles drawn normal to the strike of the small circle (transform fault) should meet at the Euler pole (Fig. 2.25a), just as the present-day circles of longitude are perpendicular to circles of latitude and converge at the geographic pole. In 1968, W. J. Morgan first used this method to locate the Euler rotation pole for the present-day plate motion between America and Africa (Fig. 2.25b). The Caribbean plate may be absorbing slow relative motion, but the absence of a well-defined seismic boundary between North and South America indicates that these plates are now moving essentially as one block. The great circles normal to transform faults in the Central Atlantic converge and intersect close to 58°N 36°W, which is an estimate of the Euler pole of recent motion between Africa and South America. The longitude of the Euler pole is determined more precisely than its latitude, the errors being ±2° and ±5°, respectively. When additional data from earthquake first motions and spreading rates are included, an Euler pole at 62°N 36°W is obtained, which is within the error of the first location.

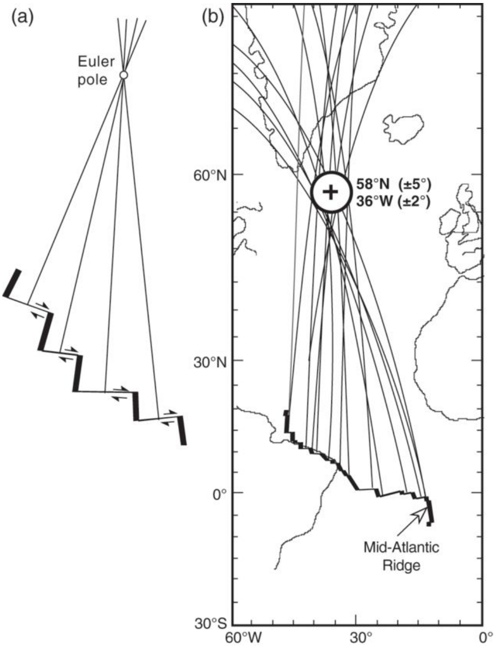


Fig. 2.25 (a) Principle of the method for locating the Euler pole of rotation between two plates where great circles normal to transform faults on the plate boundary intersect. (b) Location of the Euler pole of rotation for the motion between Africa and South America, using transform faults on the Mid-Atlantic Ridge in the Central Atlantic.

Source: W.J. Morgan, Rises, trenches, great faults, and crustal blocks, *J. Geophys. Res.* 73, 1959–1982, 1968

The “Bullard-type fit” of the African and South American coastlines (Section 2.2.2) is obtained by a rotation about a pole at 44°N 31°W. This pole reflects the average *long-term* motion between the continents. A rotation that matches a starting point with an end-point is a *finite rotation*. As the difference between the present-day and age-averaged Euler poles illustrates, a finite rotation is a mathematical formality not necessarily related to the actual motion between the plates, which may consist of a number of incremental rotations about different poles.

2.10 Forces Driving Plate Tectonic Motions

An issue that naturally arises in explaining plate tectonics is the exact mechanism that drives plate motions. The forces acting on plates may be divided into those that act on their bottom surfaces and those that act on their margins. The bottom forces arise due to relative motion between the lithospheric plate and the underlying viscous asthenosphere, the outermost layer of the mantle. Although solid, there is little doubt that on a long timescale the mantle exhibits viscous behavior, which allows it to flow and thermally driven convection to occur (see Sections 5.2 and 9.9). The dominant feature of the convection is that it involves overturn of the entire mantle on a timescale of hundreds of millions of years. According to mantle convection theory the lithosphere should be regarded as a component of the convection cycle, forming an outer thermal barrier to heat transfer. The motion vectors of lithospheric plates do not reveal directly the mantle flow pattern, but some general inferences can be drawn. The flow pattern must include the mass transport involved in moving lithosphere from a ridge to a subduction zone, which has to be balanced by return flow deep in the mantle. Interactions between the plates and a viscous substratum necessarily influence the plate motions. In order to assess the importance of these effects we need to compare them to the other forces that act on plates, especially at their boundaries (Fig. 2.26).

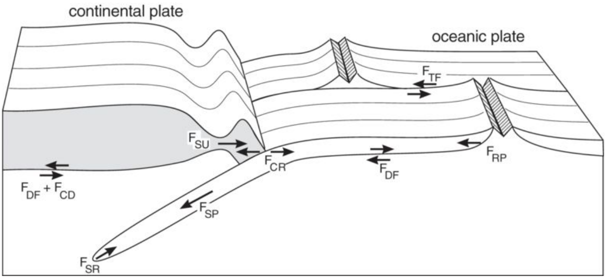


Fig. 2.26 Diagram illustrating some of the different forces acting on lithospheric plates.

Source: D. Forsyth and S. Uyeda, On the relative importance of the driving forces of plate motion, *Geophys. J. R. Astr. Soc.* 43, 163–200, 1975.

2.10.1 Forces Acting on Lithospheric Plates

Some forces acting on lithospheric plates promote motion while others resist it. Upper mantle convection could fall into either category. The flow of material beneath a plate exerts a *mantle drag force* (F_{DF}) on the base of the plate. If the convective flow is faster than plate velocities, the plates are dragged along by the flow, but if the opposite is true the mantle drag opposes the plate motion. Plate velocities are observed to be inversely related to the area of continent on the plate, which suggests that the greater lithospheric thickness results in an additional *continental drag force* (F_{CD}) on the plate. The velocity of a plate also depends on the length of its subduction zone but not on the length of its spreading ridge. This suggests that subduction forces may be more important than spreading forces. This can be evaluated by considering the forces at all three types of plate margin.

At **spreading ridges**, upwelling magma is associated with the constructive margin. It was long supposed that this process pushes the plates away from the ridge. It also elevates the ridges above the oceanic abyss, so that potential energy encourages gravitational sliding toward the trenches. Together, the two effects make up the *ridge push force* (F_{RP}).

At **transform faults**, high seismicity is evidence of interactive forces where the plates move past each other. A *transform force* (F_{TF}) can be envisioned as representing frictional resistance in the contact zone. Its magnitude may be different at a transform connecting ridge segments, where the plates are hot, than at a transform between subduction zones, where the plates are cold.

At **subduction zones**, the descending slab of lithosphere is colder and denser than the surrounding mantle. This creates a positive mass anomaly – referred to as negative buoyancy – which is accentuated by intraplate phase transitions. If the descending slab remains attached to the surface plate, a *slab pull force* (F_{SP}) ensues that pulls the slab downwards into the mantle. Transferred to the entire plate, it acts as a force toward the subduction zone. However, the subducting plate eventually sinks to depths where it approaches thermal equilibrium with the surrounding mantle, loses its negative buoyancy, and experiences a *slab resistance force* (F_{SR}) as it tries to penetrate further into the stiffer mantle.

Plate collisions result in both driving and resistive forces. The vertical pull on the descending plate may cause the bend in the lower plate to migrate away from the subduction zone, effectively drawing the upper plate toward the trench. The force on the upper plate has been termed "*trench suction*" (F_{SU}). The colliding plates also impede each other's motion and give rise to a *collision-resistance force* (F_{CR}). This force consists of separate forces due to the effects of mountains or trenches in the zone of convergence.

At **hotspots**, the transfer of mantle material to the lithosphere may result in a *hotspot force* (F_{HS}) on the plate.

In summary, the driving forces on plates are slab pull, slab suction, ridge push, and the trench pull force on the upper plate. The motion is opposed by slab resistance, collision resistance, and transform fault forces. Whether the forces between plate and mantle (mantle drag, continental drag) promote or oppose motion depends on the sense of the relative motion between the plate and the mantle. The motive force of plate tectonics is clearly a composite of these several forces. Some can be shown to be more important than others.

2.10.2 Relative Magnitudes of Forces Driving Plate Motions

In order to evaluate the relative importance of the forces, it is necessary to take into account their different directions. This is achieved by converting the forces to torques about the center of the Earth. Different mathematical analyses lead to similar general conclusions regarding the relative magnitudes of the torques. The push exerted by hotspots and the resistance at transform faults are negligible in comparison to the other forces (Fig. 2.27). The ridge push force is much smaller than the forces at a converging margin, and it is considered to be of secondary importance. Moreover, the topography of oceanic ridges is offset by transform faults. If the ridge topography were due to buoyant upwelling, the fluid mantle could not exhibit discontinuities at the faults but would bulge beyond the ends of ridge segments. Instead, sharp offsets are observed, indicating that the topography is an expression of local processes in the oceanic lithosphere. This implies that upwelling at ridges is a passive feature, with mantle material filling space created by the plates moving apart.

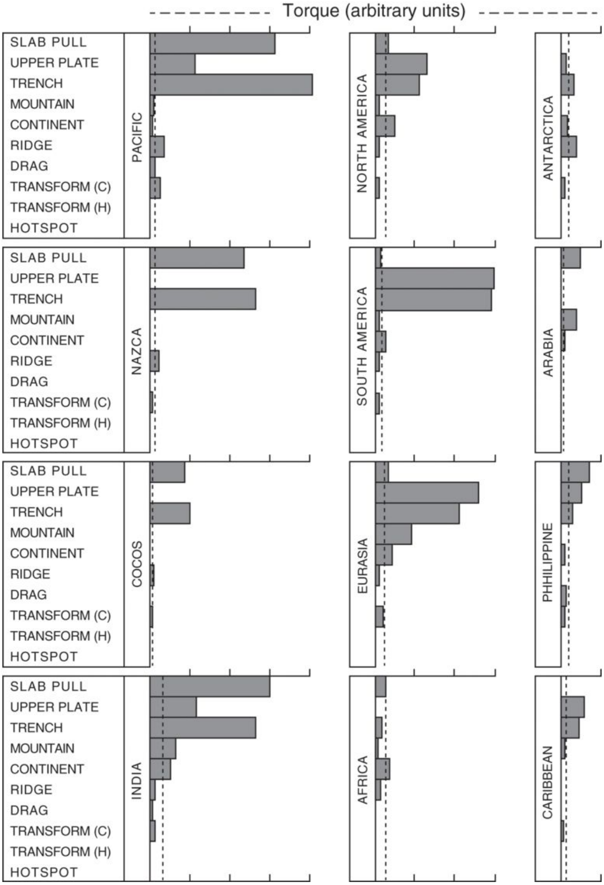


Fig. 2.27 Comparison of the magnitudes of torques acting on the 12 major lithospheric plates.

Source: W.M. Chapple and T.E. Tullis, Evaluation of the forces that drive plates, *J. Geophys. Res.* 82, 1967–1984, 1977.

The torque analysis shows that a major force driving plate motions is slab pull, which a descending slab exerts on its plate as it sinks into the mantle. The descending slab modifies local mantle circulation, which tends to draw the two plates together toward the trench that marks the subduction zone. This force – known as *slab suction* or *trench suction* – is an important driving force for plate motions. Moreover, when a subducting slab sinks into the mantle, the hinge about which it bends downward may in some cases move seawards. This is known as *hinge rollback*, or – because the trench accompanies it – as *trench rollback*. The behavior occurs primarily when oceanic plates converge and it is thought to play a role in the formation of back-arc basins.

The torque comparison reveals some other notable characteristics of the forces that affect plate motion. The resisting force due to the collision between plates is consistently smaller than the upper plate suction, and the resistance experienced by some slabs to deep mantle penetration may diminish the slab pull force. However, seismic tomography has shown that some slabs become detached from their parent plate, and may sink all the way to the core–mantle boundary. The drag forces between plate and mantle are difficult to analyze and are a topic of current research. In contrast to the early days of plate tectonic theory, where mantle drag was thought to be largely negligible, it may indeed play a more important role.

2.11 *Suggestions for Further Reading*

Introductory Level

Cox, A. and Hart, R. B. 1986. *Plate Tectonics*. Boston, MA: Blackwell Scientific.

Kearey, P., Klepeis, K. A., and Vine, F. J. 2009. *Global Tectonics*. Chichester: Wiley-Blackwell.

Molnar, P. 2015. *Plate Tectonics: A Very Short Introduction*. Oxford: Oxford University Press.

Oreskes, N. and Le Grand, H. (ed.) 2001. *Plate Tectonics: An Insider's History of the Modern Theory of the Earth*. Boulder, CO: Westview Press.

Intermediate and Advanced Level

Davies, G. F. 1999. *Dynamic Earth: Plates, Plumes and Mantle Convection*. Cambridge: Cambridge University Press.

Fowler, C. M. R. 2004. *The Solid Earth: An Introduction to Global Geophysics* (2nd ed.). Cambridge: Cambridge University Press.

Gubbins, D. 1990. *Seismology and Plate Tectonics*. Cambridge: Cambridge University Press.

Turcotte, D. L. and Schubert, G. 2002. *Geodynamics* (2nd ed.). Cambridge: Cambridge University Press.

2.12 Review Questions

1. What geological evidence is there in support of *continental drift*? What is the essential difference between older models of continental drift and the modern theory of plate tectonics?

2. What was Pangea? When and how did it form? When and how did it break up?

3. What is the Earth's crust? What is the lithosphere? How are they distinguished?

4. What are the major discontinuities in the Earth's internal structure? How are they known?

5. Distinguish between constructive, conservative, and destructive plate margins.

6. Find examples of plate margins where (a) oblique spreading and (b) oblique subduction are taking place.

7. Make a brief summary, using appropriate sketches, of geological and geophysical data from plate margins and their plate tectonic interpretations.

8. What kind of plate margin is a *continental collision* zone? How does it differ from a subduction zone?

9. Describe the Vine–Matthews–Morley hypothesis of sea-floor spreading.

10. Explain how sea-floor spreading can be used to determine the age of the oceanic crust. Where are the oldest parts of the oceans? How old are they? How does this age compare to the age of the Earth?

11. What are the names of the 12 major tectonic plates and where do the plate margins lie?

12. With the aid of a globe or map, estimate roughly a representative distance across one of the major plates. What is the ratio of this distance to the thickness of the plate? Why are the tectonic units called *plates*?

13. What is a triple junction? Explain the role of triple junctions in plate tectonics.

14. What is a hotspot? Explain how the Hawaiian hotspot provides evidence of a change in motion of the Pacific plate.

15. How may the Euler pole of relative rotation between two plates be located?

16. Explain the importance of subduction zones as the driving force of plate tectonics.

17. What do the acronyms VLBI and GPS stand for? Explain how each system functions.

2.13 Exercises

1. Summarize the geological and geophysical evidence resulting from plate tectonic activity in the following regions: (a) Iceland, (b) the Aleutian Islands, (c) Turkey, (d) the Andes, (e) the Alps?

2. Using the data in Fig. 12.31, compute the approximate spreading rates in the age interval 25–45 Ma at the oceanic ridges in the South Atlantic, South Indian, North Pacific and South Pacific Oceans.

3. Three ridges A, B, and C meet at a triple junction. Ridge A has a strike of 329° (N31W) and a spreading rate of 7.0 cm yr^{-1}; ridge B strikes at 233° (S53W) and has a spreading rate of 5.0 cm yr^{-1}. Determine the strike of ridge C and its spreading rate.

4. Three sides of a triangle on the surface of the spherical Earth measure 900 km, 1350 km, and 1450 km, respectively. Using the relationships in Box 2.1, calculate the internal angles of the triangle. If this were a plane triangle, what would the internal angles be?

5. An aircraft leaves a city at latitude λ_1 and longitude ϕ_1 and flies to a second city at latitude λ_2 and longitude ϕ_2. Using direction cosines (Box 2.2) to express the coordinates of the cities, show that the great circle distance Δ between them is given by

$$\cos\ \ = \cos\lambda_1 \cos\lambda_2 \cos(\phi_2 - \phi_1) + \sin\lambda_1 \sin\lambda_2 .$$

6. Apply the above formula to compute the great circle distances between the following pairs of cities:

(a) New York ($\lambda = 40^\circ\, 43'$ N, $\phi_1 = 74^\circ\, 1'$ W) and Madrid ($40^\circ\, 25'$ N, $3^\circ\, 43'$ W);

(b) Seattle ($\lambda = 47^\circ\, 21'$ N, $\phi_1 = 122^\circ\, 12'$ W) and Sydney

($\lambda = 33^\circ\ 52'$ S, $\phi_1 = 151^\circ\ 13'$ E);

(c) Moscow ($\lambda = 55^\circ\ 45'$ N, $\phi_1 = 37^\circ\ 35'$ E) and Paris ($\lambda = 48^\circ\ 52'$ N, $\phi_1 = 2^\circ\ 20'$ E);

(d) London ($\lambda = 51^\circ\ 30'$ N, $\phi_1 = 0^\circ\ 10'$ W) and Tokyo ($\lambda = 35^\circ\ 42'$ N, $\phi_1 = 139^\circ\ 46'$ E).

7. Calculate the heading (azimuth) of the aircraft's flight path as it leaves the first city in each pair of cities in the previous exercise.

2.14 Computational Exercises

Jupyter notebook computational exercises can be found online for this chapter at www.cambridge.org/FoG3ed (Euler pole rotation of tectonic plates).

3

Gravity and the Figure of the Earth

Preview

The shape of the Earth is determined primarily by two forces. Gravitational attraction, directed toward the center of the planet, results in an almost spherical shape. The Earth's rotation produces a centrifugal force away from the rotation axis that flattens the sphere to a rotational ellipsoid. Differences in internal mass distribution produce bumps and hollows in the ellipsoid, forming a smooth but uneven surface called the geoid. Gravity at any place acts in the vertical direction, which is everywhere perpendicular to the local geoid. We explain how the gravitational attractions of the Moon and Sun deform the Earth's free surface, creating tides in the oceans and in the solid planet. The Earth's ellipsoidal shape allows the gravitational attractions of other planets to modulate both its rotation and its orbit cyclically with periodicities of 21,000 to 405,000 years, which are correlated to long-term climatic changes.

3.1 The Earth’s Size and Shape

3.1.1 The Earth's Size

The philosophers and savants in ancient civilizations could only speculate about the nature and shape of the world they lived in. The range of possible travel was limited and only simple instruments existed. Unrelated observations might have suggested that the Earth's surface was upwardly convex. For example, the Sun's rays continue to illuminate the sky and mountain peaks after its disk has already set, departing ships appear to sink slowly over the horizon, and the Earth's shadow can be seen to be curved during partial eclipse of the Moon. However, early ideas about the heavens and the Earth were intimately bound up with concepts of philosophy, religion, and astrology. In Greek mythology the Earth was a disk-shaped region embracing the lands of the Mediterranean region and surrounded by a circular stream, Oceanus, the origin of all the rivers. In the sixth century BC the Greek philosopher Anaximander visualized the heavens as a celestial sphere that surrounded a flat Earth at its center. Pythagoras (582–507 BC) and his followers were apparently the first to speculate that the Earth was a sphere. This idea was further propounded by the influential philosopher Aristotle (384–322 BC). Although he taught the scientific principle that theory must follow fact, Aristotle is responsible for the logical device called syllogism, which can explain correct observations by apparently logical accounts that are based on false premises. His influence on scientific methodology was finally banished by the scientific revolution in the seventeenth century.

The first scientifically sound estimate of the size of the terrestrial sphere was made by Eratosthenes (275–195 BC), who was the head librarian at Alexandria, a Greek colony in Egypt during the third century BC. Eratosthenes had been told that in the city of Syene (modern Aswan) the Sun's noon rays on midsummer day shone vertically and were able to illuminate the bottoms of wells, whereas on the same day in Alexandria shadows were cast.

Using a sundial, Eratosthenes observed that at the summer solstice the Sun's rays made an angle of one-fiftieth of a circle (7.2°)

with the vertical in Alexandria (Fig. 3.1). Eratosthenes believed that Syene and Alexandria were on the same meridian. In fact they are slightly displaced; their geographic coordinates are 24° 5′N 32° 56′E and 31° 13′N 29° 55′E, respectively. Syene is actually about half a degree north of the tropic of Cancer. Eratosthenes knew that the approximate distance from Alexandria to Syene was 5000 stadia, possibly estimated by travelers from the number of days ("10 camel days") taken to travel between the two cities. From these observations Eratosthenes estimated that the circumference of the global sphere was 250,000 stadia. The Greek *stadium* was the length (about 185 m) of the U-shaped racecourse on which footraces and other athletic events were carried out. Eratosthenes' estimate of the Earth's circumference is equivalent to 46,250 km, about 15% higher than the modern value of 40,030 km.

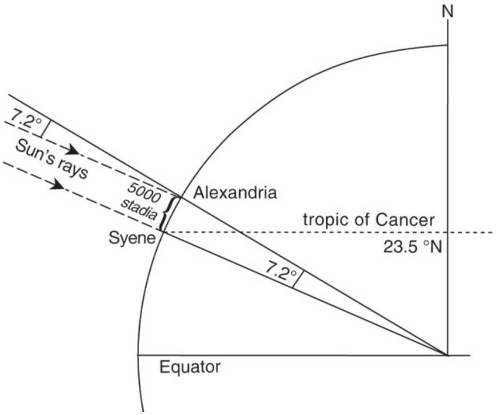


Fig. 3.1 The method used by Eratosthenes (275–195 BC) to estimate the Earth's circumference used the 7.2° difference in altitude of the Sun's rays at Alexandria and Syene, which are 5000 *stadia* apar.

Source: A.N. Strahler, *The Earth Sciences*, 681 pp., New York: Harper and Row, 1963.

Estimates of the length of one meridian degree were made in the eighth century AD during the Tang dynasty in China, and in the ninth century AD by Arab astronomers in Mesopotamia. Little progress was made in Europe until the early seventeenth century. In 1662 the Royal

Society was founded in London and in 1666 the Académie Royale des Sciences was founded in Paris. Both organizations provided support and impetus to the scientific revolution. The invention of the telescope enabled more precise geodetic surveying. In 1671 a French astronomer, Jean Picard (1620–1682), completed an accurate survey by triangulation of the length of a degree of meridian arc. From his results the Earth’s radius was calculated to be 6372 km, remarkably close to the modern value of 6371 km.

3.1.2 The Earth's Shape

In 1672 another French astronomer, Jean Richer, was sent by Louis XIV to make astronomical observations on the equatorial island of Cayenne. He found that an accurate pendulum clock, which had been adjusted in Paris precisely to beat seconds, was losing about two and a half minutes per day, i.e., its period was now too long. The error was much too large to be explained by inaccuracy of the precise instrument. The observation aroused much interest and speculation, but was only explained some 15 years later by Sir Isaac Newton in terms of his laws of universal gravitation and motion.

Newton argued that the shape of the rotating Earth should be that of an *oblate ellipsoid*; compared to a sphere, it should be somewhat flattened at the poles and should bulge outward around the equator. This inference was made on logical grounds. Assume that the Earth does not rotate and that holes could be drilled to its center along the rotation axis and along an equatorial radius (Fig. 3.2). If these holes are filled with water, the hydrostatic pressure at the center of the Earth sustains equal water columns along each radius. However, the rotation of the Earth causes a centrifugal force at the equator but has no effect on the axis of rotation. At the equator the outward centrifugal force of the rotation opposes the inward gravitational attraction and pulls the water column upward. At the same time it reduces the hydrostatic pressure produced by the water column at the Earth's center. The reduced central pressure is unable to support the height of the water column along the polar radius, which subsides. If the Earth were a hydrostatic sphere, the form of the rotating Earth should be an oblate ellipsoid of revolution. Newton assumed the Earth's density to be constant and calculated that the difference between the equatorial and polar radii is about 0.5% of the Earth's average radius, that is, approximately 32 km. This is somewhat larger than the actual flattening of the Earth, which is roughly 0.3%.

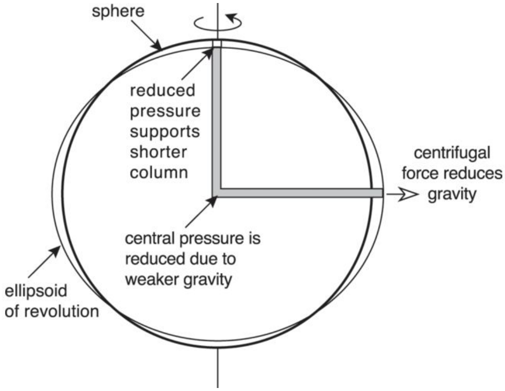

sphere
reduced pressure supports shorter column
centrifugal force reduces gravity
central pressure is reduced due to weaker gravity
ellipsoid of revolution

Fig. 3.2 Newton’s argument that the shape of the rotating Earth should be flattened at the poles and bulge at the equator was based on hydrostatic equilibrium between polar and equatorial pressure columns.

Source: A.N. Strahler, *The Earth Sciences*, 681 pp., New York: Harper and Row, 1963.

The increase in period of Richer’s pendulum could now be explained. Cayenne was close to the equator, where the larger radius placed the observer further from the center of gravitational attraction, and the increased distance from the rotational axis resulted in a stronger opposing centrifugal force. These two effects resulted in a lower value of gravity in Cayenne than in Paris, where the clock had been calibrated.

There was no direct proof of Newton’s interpretation. A corollary of his interpretation was that the degree of meridian arc should subtend a longer distance in polar regions than near the equator (Fig. 3.3). Early in the eighteenth century, French geodesists extended the standard meridian from border to border of the country and found a puzzling result. In contrast to the prediction of Newton, the degree of meridian arc *decreased* northward. The French interpretation was that the Earth’s shape was a *prolate* ellipsoid, elongated at the poles and narrowed at the equator, like the shape of a rugby football. A major scientific controversy arose between the “flatteners” and the “elongators.”

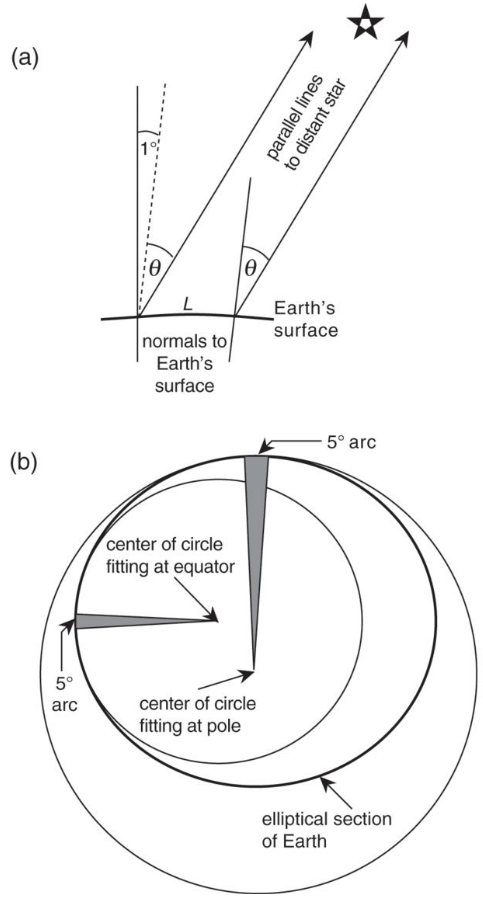

(a)
1°
θ
θ
L
parallel lines to distant star
Earth's surface
normals to Earth's surface
(b)
5° arc
center of circle fitting at equator
5° arc
center of circle fitting at pole
elliptical section of Earth

Fig. 3.3 (a) The length of a degree of meridian arc is found by measuring the distance between two points that lie one degree apart on the same meridian. (b) The larger radius of curvature at the flattened poles gives a longer arc distance than is found at the equator, where the radius of curvature is smaller.

Source: A.N. Strahler, *The Earth Sciences*, 681 pp., New York: Harper and Row, 1963.

To determine whether the Earth's shape is oblate or prolate, the Académie Royale des Sciences sponsored two scientific expeditions. In 1736–1737 a team of scientists measured the length of a degree of meridian arc in Lapland, near the Arctic Circle. They found a length appreciably longer than the meridian degree measured by Picard near Paris. From 1735 to 1743 a second party of scientists measured the length of more than three degrees of meridian arc in Peru, near the equator. Their results showed that the equatorial degree of latitude was shorter than the meridian degree in Paris. Both parties confirmed convincingly the prediction of Newton that the Earth’s shape is that of an oblate ellipsoid.

The ellipsoidal shape of the Earth resulting from its rotation has important consequences, not only for the variation with latitude of gravity on the Earth’s surface, but also for the Earth’s rate of rotation and the orientation of its rotational axis. These are modified by torques that arise from the gravitational attractions of the Sun, Moon, and planets on the ellipsoidal shape.

3.2 Gravitation

3.2.1 The Law of Universal Gravitation

Sir Isaac Newton (1642–1727) was born in the same year that Galileo died. Unlike Galileo, who relished debate, Newton was a retiring person and avoided confrontation. This is apparent in a letter written in 1675 to his colleague Robert Hooke, famous for his experiments on elasticity. In this letter Newton made the famous disclaimer "if I have seen further [than you and Descartes], it is by standing upon the shoulders of Giants." In modern terms Newton would be regarded as a theoretical physicist. He had an outstanding ability to synthesize experimental results and incorporate them into his own theories. Faced with the need for a more powerful technique of mathematical analysis than existed at the time, he invented differential and integral calculus, for which he is credited equally with Gottfried Wilhelm von Leibnitz (1646–1716), who discovered the same method independently. Newton was able to resolve many issues by formulating logical thought experiments; an example is his prediction that the shape of the Earth is an oblate ellipsoid. He was one of the most outstanding synthesizers of observations in scientific history, which is implicit in his letter to Hooke. His three-volume book, *Philosophiae Naturalis Principia Mathematica*, published in 1687, ranks as the greatest of all scientific texts. The first volume of the *Principia* contains Newton's famous *Laws of Motion*, the third volume handles the *Law of Universal Gravitation*.

The first two laws of motion are generalizations from Galileo's results. As a corollary Newton applied his laws of motion to demonstrate that forces must be added as vectors and showed how to do this geometrically with a parallelogram. The second law of motion states that the rate of change of momentum of a mass is proportional to the force acting upon it and takes place in the direction of the force. For the case of constant mass, this law serves as the definition of force (**F**) in terms of the acceleration (**a**) given to a mass (*m*):

$$\mathbf{F} = m\mathbf{a}.$$

(3.1)

The unit of force in the SI system of units is the newton (N). It is defined as the force that gives a mass of one kilogram (1 kg) an acceleration of $1\ \mathrm{m\ s^{-2}}$.

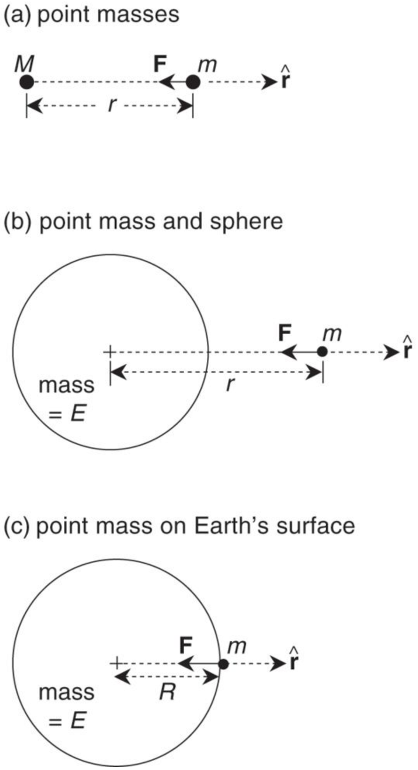

(a) point masses
M
F
m
r
r̂
(b) point mass and sphere
F
m
r̂
mass
= E
r
(c) point mass on Earth's surface
F
m
r̂
mass
= E
R

Fig. 3.4 Geometries for the gravitational attraction on (a) two point masses, (b) a point mass outside a sphere, and (c) a point mass on the surface of a sphere.

His celebrated observation of a falling apple may be a legend, but Newton's genius lay in recognizing that the type of gravitational field that caused the apple to fall was the same type that served to hold the Moon in its orbit around the Earth, the planets in their orbits around the Sun, and that acted between minute particles characterized only by their masses. Newton used Kepler's empirical third law (see Section 1.3 and Eq. 1.6) to deduce that the force of attraction between a planet and the Sun varied with the “quantities of solid matter that they contain” (i.e., their masses) and with the inverse square of the distance between them. Applying this law to two particles or point masses m and M separated by a distance r (Fig. 3.4a), we get for the gravitational attraction **F** exerted by M on m

$$\mathbf{F} = -G\frac{mM}{r^2}\hat{\mathbf{r}}. \tag{3.2}$$

In this equation, $\hat{\mathbf{r}}$ is a unit vector in the direction of increase of coordinate r, directed away from the center of reference at the mass M. The negative sign in the equation indicates that the force **F** acts in the opposite direction, toward the attracting mass M. The constant G is the constant of universal gravitation. An important aspect of Eq. (3.2) is that the roles of m and M may be interchanged, meaning that it also describes the attraction of mass M by mass m.

3.2.1.1 The Gravitational Constant, G

There was no way to determine the gravitational constant experimentally during Newton’s lifetime. The method to be followed was evident, namely to determine the force between two masses in a laboratory experiment. However, seventeenth-century technology was not yet up to this task. Experimental determination of G was extremely difficult. It was first measured more than a century after the publication of *Principia* by Lord Charles Cavendish (1731–1810). From a set of painstaking measurements of the force of attraction between two spheres of lead, Cavendish in 1798 determined the value of G to be $6.754 \times 10^{-11}\ \mathrm{m^3\,kg^{-1}\,s^{-2}}$. The CODATA-2014 recommended values of the fundamental physical constants (Mohr et al., 2016) gives the latest estimate of G as $6.67408 \times 10^{-11}\ \mathrm{m^3\,kg^{-1}\,s^{-2}}$.

The value of G must be determined directly by laboratory measurement, and this is technically very difficult, as evident in the amount of scatter in the mean values of individual determinations (Box 3.1). The best estimate for G has a comparatively large relative uncertainty of 4.7 x 10^{-5} (47 parts per million). As a result, the gravitational constant is the most poorly known of the major physical constants, most of which have relative uncertainties on the order of 10^{-7} to 10^{-11}. By contrast, the product of the gravitational constant and the Earth's mass, GE – called the *geocentric gravitational constant* – can be derived by means of highly accurate laser tracking of artificial satellites. Consequently, GE is known with a precision of around 2×10^{-9}, similar to the uncertainties of the major physical constants. The current best estimate for the value of GE is $3{,}986{,}004.418 \pm 0.008 \times 10^{8}$ m^3 s^{-2}.

Box 3.1 Determining the Gravitational Constant

Gravitation is one of four fundamental forces of physics, along with electromagnetic force and the strong and weak nuclear forces. It is vastly weaker than the other three. As a result, the constant G cannot be determined indirectly from other physical measurements but must be determined independently by laboratory measurement. This is technically very difficult, partly because an experimental measurement can be affected by external gravitational attractions. Although different physical principles have been employed, most experiments have been conducted with torsion balances. The experiments attempt to compensate for disturbing effects; for example, by housing the sensors in vacuum to obviate air resistance, or by measuring at cryogenic temperature to remove thermal effects.

The errors in 14 determinations made in the past 20 years (Fig. B3.1.1) are very small in some experiments, but the mean values obtained by different laboratories are quite scattered. Even precise measurements (with small standard errors) are quite distinct from each other (their error bars do not overlap).

This lack of agreement between individual determinations is not understood.

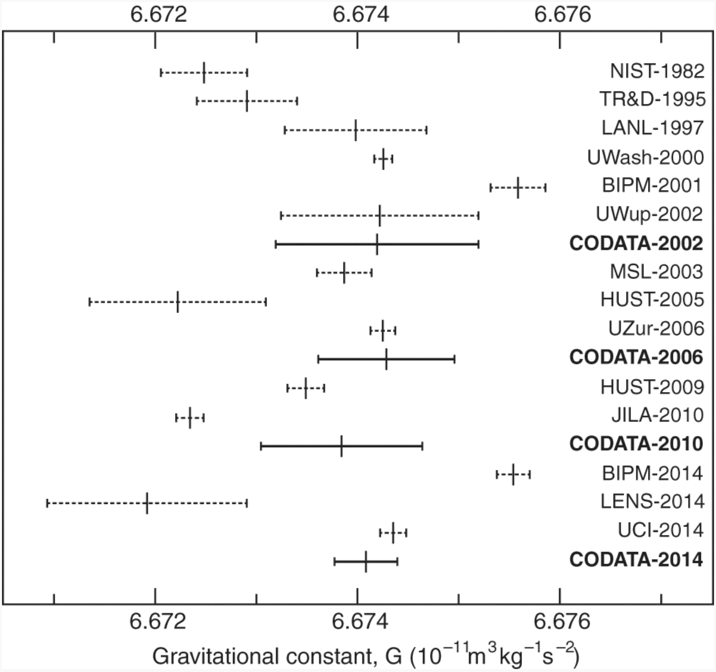


Fig. B3.1.1 Measured values of the gravitational constant in the past 20 years. Letters on the right identify the source and the year of the measurement. CODATA values are optimized averages for all preceding measurements

The current best estimate for G is $6.67408 \pm 0.00031 \times 10_{-}^{11}$ kg$_{-}^{1}$ m^3 s$_{-}^{2}$ (CODATA-2014: Mohr et al., 2016).

3.2.1.2 Potential Energy and Work

The law of conservation of energy means that the total energy of a closed system is constant. Two forms of energy need be considered here. The first is the potential energy, which an object has by virtue of its position relative to the origin of a force. The second is the work done against the action of the force during a change in position.

For example, when Newton's apple is on the tree it has a higher potential energy than when it lies on the ground. It falls because of the downward force of gravity and loses potential energy in doing so. To compute the change in potential energy we need to raise the apple to its original position. This requires that we apply a force equal and opposite to the gravitational attraction on the apple, and, because this force must be moved through the distance the apple fell, we have to expend energy in the form of work. If the original height of the apple above ground level was h and the value of the force exerted by gravity on the apple is F, the force we must apply to put it back is $-F$. Assuming that F is constant through the short distance of its fall, the work expended is $-Fh$. This is the increase in potential energy of the apple when it is on the tree.

More generally, if the constant force F moves through a small distance dr in the same direction as the force, the work done is $\mathrm{d}W = F\,\mathrm{d}r$ and the change in potential energy $\mathrm{d}E_\mathrm{p}$ is given by

$$\mathrm{d}E_\mathrm{p} = -\mathrm{d}W = -F\mathrm{d}r. \tag{3.3}$$

In the more general case we have to consider motions and forces that have components along three orthogonal axes. The displacement dr and the force F no longer need to be parallel to each other. We have to treat F and dr as vectors. In Cartesian coordinates the displacement vector d**r** has components (dx, dy, dz) and the force **F** has components (F_x, F_y, F_z) along each of the respective axes. The work done by the x-component of the force when it is displaced along the x-axis is $F_x\,\mathrm{d}x$, and there are similar expressions for the displacements along the other axes. The change in potential energy $\mathrm{d}E_\mathrm{p}$ is now given by

$$\mathrm{d}E_\mathrm{p} = -\mathrm{d}W = -(F_x\mathrm{d}x + F_y\mathrm{d}y + F_z\mathrm{d}z). \tag{3.4}$$

The expression in brackets is called the *scalar product* – written **F**·d**r** –

of the vectors **F** and d**r**. Its scalar value is equal to F dr cosθ, where θ is the angle between the vectors.

3.2.2 Gravitational Acceleration

In physics the *field* of a force is often more important than the absolute magnitude of the force. The field is defined as the force exerted on a material unit. For example, the electrical field of a charged body at a certain position is the force it exerts on a unit of electrical charge at that location. The *gravitational field* in the vicinity of an attracting mass is the force it exerts on a unit mass. Equation (3.1) shows that this is equivalent to the acceleration vector.

In geophysical applications we are concerned with accelerations rather than forces. By comparing Eq. (3.1) and Eq. (3.2) we get the gravitational acceleration $\mathbf{a}_G$ of the mass m due to the attraction of the mass M:

$$\mathbf{a}_G = -G\frac{M}{r^2}\hat{\mathbf{r}}. \tag{3.5}$$

The SI unit of acceleration is m s^{-2}; this unit is rather large for practical use in geophysics. In the now superseded c.g.s. system the unit of acceleration was the cm s^{-2}, which is called a Gal in recognition of the contributions of Galileo. The small changes in the acceleration of gravity caused by geological structures are measured in thousandths of this unit, i.e., in milligal (mGal). Until recently, gravity anomalies due to geological structures were surveyed with field instruments accurate to about one-tenth of a milligal, which was called a *gravity unit*. Modern instruments are capable of measuring gravity differences to a millionth of a Gal, or microgal (μGal), which is now a common unit in gravity investigations. The most sensitive gravity meters can detect gravity differences on the scale of 10^{-9} Gal, which is called a nanogal (nGal). This is equivalent to one-trillionth (10^{-12}) of the value of gravity at the Earth's surface, which averages approximately 9.81 m s^{-2}.

3.2.2.1 Gravitational Potential

The gravitational potential is the potential energy of a unit mass in a field of gravitational attraction. Let the potential be denoted by the symbol U_G. The potential energy E_p of a mass m in a gravitational field is thus equal to ($m\ U_G$). Thus, a change in potential energy (dE_p) is equal to ($m\ dU_G$). Equation (3.3) becomes, by using Eq. (3.1),

$$m\mathrm{d}U_G = -F\mathrm{d}r = -ma_G\mathrm{d}r. \tag{3.6}$$

Rearranging and taking directions into account, we get the gravitational acceleration

$$\mathbf{a}_\mathrm{G} = -\frac{\mathrm{d}U_\mathrm{G}}{\mathrm{d}r}\hat{\mathbf{r}}. \tag{3.7}$$

In general, the acceleration is a three-dimensional vector. If we are using Cartesian coordinates (x, y, z), the acceleration will have components (a_x, a_y, a_z). These may be computed by separately calculating the derivatives of the potential with respect to x, y, and z:

$$a_x = -\frac{\partial U_\mathrm{G}}{\partial x} \qquad a_y = -\frac{\partial U_\mathrm{G}}{\partial y} \qquad a_z = -\frac{\partial U_\mathrm{G}}{\partial z}. \tag{3.8}$$

Equating Eqs. (3.3) and (3.7) gives the gravitational potential of a point mass M:

$$\frac{\mathrm{d}U_\mathrm{G}}{\mathrm{d}r} = G\frac{M}{r^2}, \tag{3.9}$$

the solution of which is

$$U_G = -G\frac{M}{r}.$$
(3.10)

3.2.2.2 Acceleration and Potential of a Distribution of Mass

Until now, we have considered only the gravitational acceleration and potential of point masses. A solid body may be considered to be composed of numerous small particles, each of which exerts a gravitational attraction at an external point P (Fig. 3.5a). To calculate the gravitational acceleration of the object at point P we must form a vector sum of the contributions of the individual discrete particles. Each contribution has a different direction. Assuming m_i to be the mass of the particle at distance r_i from P, this gives an expression like

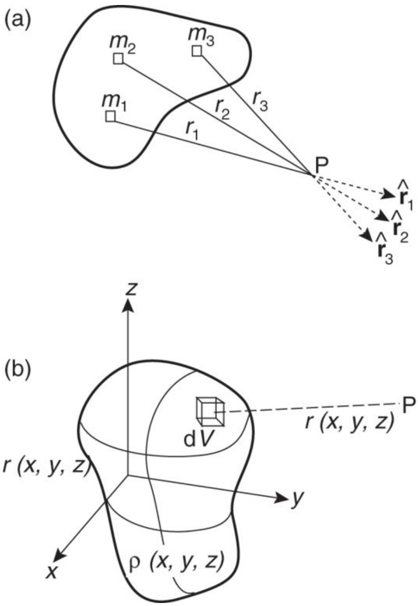


Fig. 3.5 (a) Each small particle of a solid body exerts a gravitational attraction in a different direction at an external point P. (b) Computation of the gravitational potential of a continuous mass distribution.

$$\mathbf{a}_G = -G\frac{m_1}{r_1^2}\hat{\mathbf{r}}_1 - G\frac{m_2}{r_2^2}\hat{\mathbf{r}}_2 - G\frac{m_3}{r_3^2}\hat{\mathbf{r}}_3 - \cdots.$$

(3.11)

Depending on the shape of the solid, this vector sum can be quite complicated.

An alternative solution to the problem is by first calculating the gravitational potential, and then differentiating it as in Eq. (3.5) to get the acceleration. The expression for the potential at P is

$$U_G = -G\frac{m_1}{r_1} - G\frac{m_2}{r_2} - G\frac{m_3}{r_3} - \cdots. \tag{3.12}$$

This is a scalar sum, which is usually more simple to calculate than a vector sum.

More commonly, the object is not represented as an assemblage of discrete particles but by a continuous mass distribution. However, we can subdivide the volume into discrete elements; if the density of the matter in each volume is known, the mass of the small element can be calculated and its contribution to the potential at the external point P can be determined. By integrating over the volume of the body, its gravitational potential at P can be calculated. At a point in the body with coordinates (x, y, z) let the density be $\rho(x, y, z)$ and let its distance from P be $r(x, y, z)$ as in Fig. 3.5b. The gravitational potential of the body at P is

$$U_G = -G\int_x\int_y\int_z \frac{\rho(x,y,z)}{r(x,y,z)}\,\mathrm{d}x\mathrm{d}y\mathrm{d}z. \tag{3.13}$$

The integration gives the gravitational potential and acceleration at points inside and outside a hollow or homogeneous solid sphere. The values outside a sphere at distance r from its center are the same as if the entire mass E of the sphere were concentrated at its center (Fig. 3.4b):

$$U_G = -G\frac{E}{r},$$
(3.14)

$$\mathbf{a}_G = -G\frac{E}{r^2}\hat{\mathbf{r}}.$$
(3.15)

3.2.2.3 Mass and Mean Density of the Earth

Equations (3.14) and (3.15) are valid everywhere outside a sphere, including on its surface where the distance from the center of mass is equal to the mean radius R (Fig. 3.4c). If we regard the Earth to a first approximation as a sphere with mass E and radius R, we can estimate the Earth's mass by rewriting Eq. (3.15) as a scalar equation in the form

$$E = \frac{R^2 a_G}{G}.$$
(3.16)

The gravitational acceleration at the surface of the Earth, which is not perfectly spherical, is about 9.81 $\mathrm{m\,s^{-2}}$. The Earth's mean radius is 6371 km, and the gravitational constant is $6.674 \times 10^{-11}\ \mathrm{m^3\,kg^{-1}\,s^{-2}}$. This gives the mass of the Earth as approximately 5.973×10^{24} kg. This large number is more difficult to appreciate than the mean density of the Earth, which may be calculated by dividing the Earth's mass by its volume ($4/3\,\pi R^3$). A mean density of 5514 $\mathrm{kg\,m^{-3}}$ is obtained, which is about double the density of crustal rocks. This suggests that the Earth's interior is not homogeneous, and that density must, in general, increase with depth in the Earth.

3.2.3 The Equipotential Surface

An equipotential surface is one on which the potential is constant. For a sphere of given mass the gravitational potential (Eq. 3.14) varies only with the distance r from its center. A certain value of the potential, say U_1, is realized at a constant radial distance r_1. Thus, the equipotential surface on which the potential has the value U_1 is a sphere with radius r_1; a different equipotential surface U_2 is the sphere with radius r_2. The equipotential surfaces of the original spherical mass form a set of concentric spheres (Fig. 3.6a), one of which (e.g., U_0) coincides with the surface of the spherical mass. This particular equipotential surface describes the *figure* of the spherical mass.

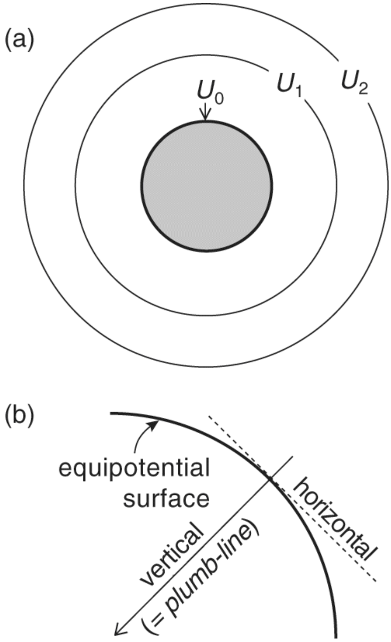


Fig. 3.6 (a) Equipotential surfaces of a spherical mass form a set of concentric spheres. (b) The normal to the equipotential surface defines the vertical direction; the tangential plane defines the horizontal.

By definition, no change in potential takes place (and no work is done) in moving from one point to another on an equipotential surface. The work done by a force F in a displacement dr is F dr cosθ, which is zero when cosθ is zero, that is, when the angle θ between the displacement and the force is 90°. If no work is done in a motion along a gravitational equipotential surface, the force and acceleration of the gravitational field must act perpendicular to the surface. This normal to the equipotential surface defines the *vertical*, or plumb-line, direction (Fig. 3.6b). The plane tangential to the equipotential surface at a point defines the *horizontal* at that point.

When the distribution of masses inside the body is not spherically symmetric, the equipotential surfaces are no longer spherical. Nevertheless, moving on this more complicated equipotential surface does not require any work, and so the definitions of vertical and horizontal remain meaningful.

3.3 The Earth’s Rotation

3.3.1 Introduction

The Earth behaves as an elastic body and deforms in response to the forces generated by its rotation, becoming slightly flattened at the poles with a compensating bulge at the equator. The gravitational attractions of the Sun, Moon, and planets on the spinning, flattened Earth cause changes in its rate of rotation, in the orientation of the rotation axis, and in the shape of the Earth’s orbit around the Sun. Even without extra-terrestrial influences the Earth reacts to tiny displacements of the rotation axis from its average position by acquiring a small, unsteady wobble. These perturbations reflect a balance between gravitation and the forces that originate in the Earth’s rotational dynamics.

3.3.2 Centripetal and Centrifugal Acceleration

Newton's first law of motion states that every object continues in its state of rest or of uniform motion in a straight line unless compelled to change that state by forces acting on it. The continuation of a state of motion is by virtue of the *inertia* of the body. A framework in which this law is valid is called an *inertial system*. For example, when we are traveling in a car at constant speed, we feel no disturbing forces; reference axes fixed to the moving vehicle form an inertial frame. If traffic conditions compel the driver to apply the brakes, we experience decelerating forces; if the car goes around a corner, even at constant speed, we sense sideways forces toward the outside of the corner. In these situations the moving car is being forced to change its state of uniform rectilinear motion and reference axes fixed to the car form a *non-inertial system*.

Motion in a circle implies that a force is active that continually changes the state of rectilinear motion. Newton recognized that the force was directed inwards, toward the center of the circle, and named it the *centripetal* (meaning "center-seeking") force. He cited the example of a stone being whirled about in a sling. The inward centripetal force exerted on the stone by the sling holds it in a circular path. If the sling is released, the restraint of the centripetal force is removed and the inertia of the stone causes it to continue its motion at the point of release. No longer under the influence of the restraining force, the stone flies off in a straight line. Arguing that the curved path of a projectile near the surface of the Earth was due to the effect of gravity, which caused it constantly to fall toward the Earth, Newton postulated that, if the speed of the projectile were exactly right, it might never quite reach the Earth's surface. If the projectile fell toward the center of the Earth at the same rate as the curved surface of the Earth fell away from it, the projectile would go into orbit around the Earth. Newton suggested that the Moon was held in orbit around the Earth by just such a centripetal force, which originated in the gravitational attraction of the Earth. Likewise, he visualized that a centripetal force due to gravitational attraction restrained the planets in their circular orbits about the Sun.

The passenger in a car going round a corner experiences a tendency to be flung outwards. He is restrained in position by the frame of the vehicle, which supplies the necessary centripetal acceleration to enable the passenger to go round the curve in the car. The inertia of the passenger's body causes it to continue in a straight line and pushes him outwards against the side of the vehicle. This outward force is called the *centrifugal* force. It arises because the car does not represent an inertial reference frame. An observer outside the car in a fixed (inertial) coordinate system would note that the car and passenger are constantly changing direction as they round the corner. The centrifugal force feels real enough to the passenger in the car, but it is called a pseudo-force, or *inertial force*. In contrast to the centripetal force resulting from the car's restraint, the centrifugal force does not have a physical origin, but exists only because it is being observed in a non-inertial reference frame.

3.3.2.1 Centripetal Acceleration

The mathematical form of the centripetal acceleration for circular motion with constant angular velocity ω (in radians per second) about a point can be derived as follows. Define orthogonal Cartesian axes x and y relative to the center of the circle as in Fig. 3.7a. The linear velocity $\mathbf{v}$ at any point where the radius vector makes an angle $\theta = (\omega t)$ with the x-axis has components

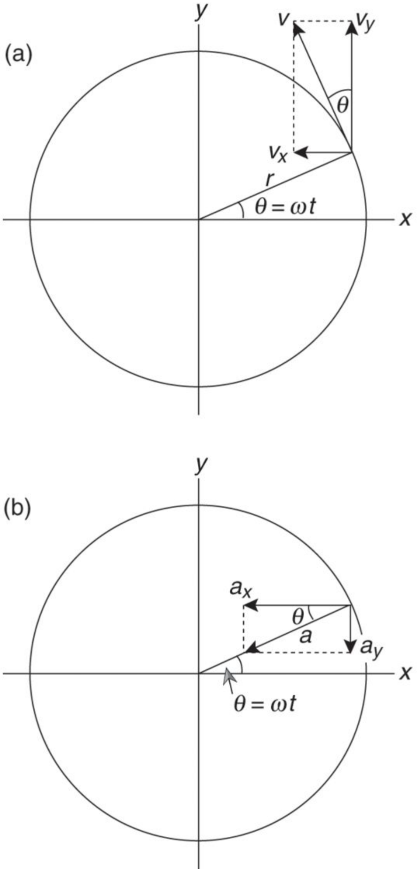


Fig. 3.7 (a) Components v_x and v_y of the linear velocity v where the radius makes an angle $\theta = (\omega t)$ with the x-axis, and (b) the components a_x and a_y of the centripetal acceleration, which is directed radially inward.

$$\begin{aligned} v_x &= -v\sin(\omega t) = -r\omega\sin(\omega t) \\ v_y &= v\cos(\omega t) = r\omega\cos(\omega t). \end{aligned} \tag{3.17}$$

The x- and y-components of the acceleration are obtained by differentiating the velocity components with respect to time. This gives

$$\begin{aligned} a_x &= -\omega v\cos(\omega t) = -r\omega^2\cos(\omega t) \\ a_y &= -\omega v\sin(\omega t) = -r\omega^2\sin(\omega t). \end{aligned}$$

(3.18)

These are the components of the centripetal acceleration, which is directed radially inwards and has the magnitude $\omega^2 r$ (Fig. 3.7b).

3.3.2.2 Centrifugal Acceleration and Potential

In handling the variation of gravity on the Earth's surface we must operate in a non-inertial reference frame attached to the rotating Earth. Viewed from a fixed, external inertial frame, a stationary mass moves in a circle about the Earth's rotation axis with the same rotational speed as the Earth. However, within a rotating reference frame attached to the Earth, the mass is stationary. It experiences a centrifugal acceleration (a_c) that is exactly equal and opposite to the centripetal acceleration, and which can be written

$$a_c = \omega^2 r = \frac{v^2}{r}.$$

(3.19)

The centrifugal acceleration is not a centrally oriented acceleration like gravitation, but instead is defined relative to an axis of rotation. Nevertheless, potential energy is associated with the rotation and it is possible to define a *centrifugal potential*. Consider a point rotating with the Earth at a distance r from its center (Fig. 3.8). The angle θ between the radius to the point and the axis of rotation is called the colatitude; it is the angular complement of the latitude λ. The distance of the point from the rotational axis is x ($= r \sin\theta$), and the centrifugal acceleration is $\omega^2 x$ outwards in the direction of increasing x. The centrifugal potential U_c is defined such that

$$\mathbf{a}_c = -\frac{\partial U_c}{\partial x}\hat{\mathbf{x}} = (\omega^2 x)\hat{\mathbf{x}},$$

(3.20)

where $\hat{\mathbf{x}}$ is the outward unit vector. On integrating, we obtain

$$U_c = -\frac{1}{2}\omega^2 x^2 = -\frac{1}{2}\omega^2 r^2 \cos^2\lambda$$
$$= -\frac{1}{2}\omega^2 r^2 \sin^2\theta.$$
(3.21)

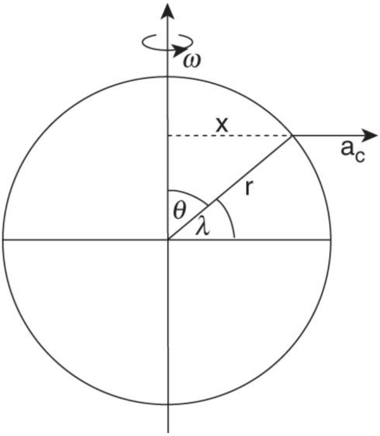


Fig. 3.8 The outwardly directed centrifugal acceleration a_c at latitude λ on a sphere rotating at angular velocity ω.

3.3.2.3 Kepler's Third Law of Planetary Motion

By comparing the centripetal acceleration of a planet about the Sun with the gravitational acceleration of the Sun, the third of Kepler's laws of planetary motion can be explained (see Section 1.3). Let S be the mass of the Sun, r_p the distance of a planet from the Sun, and T_p the period of orbital rotation of the planet around the Sun. Equating the gravitational and centripetal accelerations gives

$$G\frac{S}{r_\mathrm{p}^2} = \omega_\mathrm{p}^2 r_\mathrm{p} = \left(\frac{2\pi}{T_\mathrm{p}}\right)^2 r_\mathrm{p}.$$
(3.22)

Rearranging this equation, we get Kepler's third law of planetary motion, which states that the square of the period of the planet is proportional to the cube of the radius of its orbit, or:

$$\frac{r_{\mathrm{P}}^3}{T_{\mathrm{P}}^2} = \frac{GS}{4\pi^2} = \text{constant.}$$
(3.23)

3.3.2.4 Verification of the Inverse Square Law of Gravitation

Newton realized that the centripetal acceleration of the Moon in its orbit was supplied by the gravitational attraction of the Earth, and tried to use this knowledge to confirm the inverse square dependency on distance in his law of gravitation. The sidereal period (T_{L}) of the Moon about the Earth, a sidereal month, is equal to 27.3 days. Let the corresponding angular rate of rotation be $\omega_{\mathrm{L}} = 2\pi/T_{\mathrm{L}}$. We can equate the gravitational acceleration of the Earth at the Moon with the centripetal acceleration due to ω_{L}:

$$G\frac{E}{r_{\mathrm{L}}^2} = \omega_{\mathrm{L}}^2 r_{\mathrm{L}}.$$
(3.24)

This equation can be rearranged as follows:

$$\left(G\frac{E}{R^2}\right)\left(\frac{R}{r_{\mathrm{L}}}\right)^2 = \omega_{\mathrm{L}}^2 R\left(\frac{r_{\mathrm{L}}}{R}\right).$$
(3.25)

Comparison with Eq. (3.15) shows that the first quantity in parentheses is the mean gravitational acceleration on the Earth's surface, a_{G}. Therefore, we can write

$$a_{\mathrm{G}} = G\frac{E}{R^2} = \omega_{\mathrm{L}}^2 R\left(\frac{r_{\mathrm{L}}}{R}\right)^3.$$
(3.26)

In Newton's time little was known about the physical dimensions of the Earth. The distance of the Moon was estimated by using the method of parallax (see Section 1.2) to be approximately 60 times the radius of the Earth and its sidereal period was known to be 27.3 days. At first Newton used the accepted value of 5500 km for the Earth's radius. This gave a value of only 8.4 m s^{-2} for gravity, well below the known value of 9.8 m s^{-2}. However, in 1671 Picard determined the Earth's radius to be 6372 km. With this value, the inverse square character of Newton's law of gravitation was confirmed.

3.3.3 Coriolis and Eötvös Accelerations

The centrifugal acceleration due to the Earth's rotation affects moving objects by causing additional accelerations related to the velocity at which they are moving. At latitude λ the distance d of a point on the Earth's surface from the rotational axis is equal to $R\cos\lambda$, and the rotational spin ω translates to an eastwards linear velocity v equal to $\omega R\cos\lambda$ (Fig. 3.9a). Consider an object (e.g., a vehicle or projectile) that is moving at velocity v across the Earth's surface. In general v has a northward component v_N and an eastward component v_E. The eastward velocity is added to the linear velocity of the rotation. As a result, the centrifugal acceleration increases by an amount Δa_c, which can be obtained by differentiating a_c in Eq. (3.19) with respect to ω:

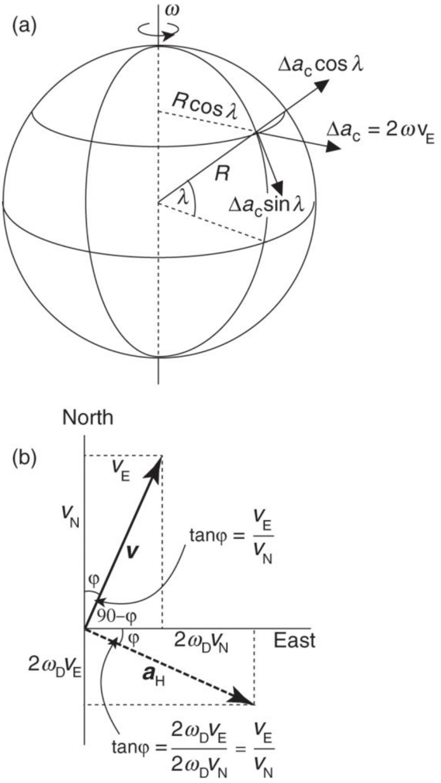

(a)
ω
$\Delta a_c \cos\lambda$
$R\cos\lambda$
$\Delta a_c = 2\omega v_E$
R
λ
$\Delta a_c \sin\lambda$
North
(b)
V_E
V_N
$\boldsymbol{v}$
$\tan\varphi = \frac{V_E}{V_N}$
φ
$90-\varphi$
φ
$2\omega_D V_N$
East
$2\omega_D V_E$
$\boldsymbol{a}_H$
$\tan\varphi = \frac{2\omega_D V_E}{2\omega_D V_N} = \frac{V_E}{V_N}$

Fig. 3.9 (a) Vertical and horizontal components of the additional centrifugal acceleration Δa_c at latitude λ due to eastward velocity $\mathbf{v}_E$

on a rotating Earth. (b) Relationship between the components of velocity **v** in the horizontal plane and the components of the Coriolis acceleration a_H.

$$\Delta a_c = 2\omega(R\cos\lambda)\Delta\omega = 2\omega v_E. \tag{3.27}$$

Eötvös acceleration: The extra centrifugal acceleration Δa_c can be resolved into a vertical component and a horizontal component (Fig. 3.9a). The vertical component ($\Delta a_c \cos\lambda = 2\omega\, v_E \cos\lambda$) acts upward, opposite to gravity. It is called the *Eötvös acceleration*. Its effect is to *decrease* the measured gravity by an amount that depends on the eastern component of the object's speed, v_E. If the moving object has a westward component of velocity, the Eötvös acceleration acts downward and *increases* the measured gravity. It affects gravity measurements that are made on a moving platform (for example, on a research ship or in an airplane). The measurements must be corrected to compensate for the Eötvös effect. A ship sailing eastward at 10 km h^{-1} at latitude 45° requires an Eötvös correction of 28.6 mGal; in an airplane flying eastward at 300 km h^{-1} the correction is 856 mGal. These corrections are far greater than the sizes of many important gravity anomalies. However, the Eötvös correction can be made satisfactorily in marine gravity surveys, and recent technical advances make it feasible in airborne gravimetry (Section 4.1.4.1).

Coriolis acceleration: The horizontal component a_N of the acceleration caused by the eastwards horizontal velocity v_E acts southward and has the value

$$a_N = -(2\omega\sin\lambda)v_E = -2\omega_V v_E. \tag{3.28}$$

Note that this acceleration depends on the vertical component ω_V ($= \omega \sin \lambda$) of the Earth's rotation ω at latitude λ. For an eastward-moving object in the northern hemisphere a_N acts to the south, as indicated by the negative sign (conversely, if the object is moving westward, the acceleration is northward). This component of the Coriolis acceleration arises from the eastward component of the object's motion. A further component of the Coriolis acceleration derives from the northward component of the object's motion, and can be derived as follows.

From Eqs. (1.1) and (1.2) the angular momentum h of a mass m at latitude λ on the rotating Earth is

$$h = m\omega(R\cos\lambda)^2. \tag{3.29}$$

Conservation of angular momentum requires that h is constant in time.

$$\frac{\partial h}{\partial t} = mR^2\cos^2\lambda\frac{\partial\omega}{\partial t} + m\omega R^2(-2\cos\lambda\sin\lambda)\frac{\partial\lambda}{\partial t} = 0. \tag{3.30}$$

Rearranging and simplifying, we get

$$(R\cos\lambda)\frac{\partial\omega}{\partial t} = 2\omega\sin\lambda\left(R\frac{\partial\lambda}{\partial t}\right). \tag{3.31}$$

The expression in brackets on the right of the equation involves the rate of increase of latitude and is the northward velocity component v_N. The expression on the left is an acceleration, a_E, which is the eastward component of the Coriolis acceleration:

$$a_E = (2\omega\sin\lambda)v_N. \tag{3.32}$$

The north and east components of the Coriolis acceleration are therefore:

$$a_N = -2(\omega\sin\lambda)v_E = -2\omega_V v_E$$
$$a_E = 2(\omega\sin\lambda)v_N = 2\omega_V v_N.$$
(3.33)

These components are plotted in Fig. 3.9b for an object moving with horizontal velocity v in the northern hemisphere. If the heading of the object is φ relative to north, the direction of the Coriolis acceleration, a_H, makes the same angle φ with the eastwards direction (the tangent of the angle in each case is v_E/v_N). The Coriolis acceleration therefore acts perpendicular to the velocity. In the *northern* hemisphere it acts horizontally to the *right* of the direction of motion. In the *southern* hemisphere the sense of this acceleration is reversed; it acts to the *left* of the direction of motion.

The Coriolis acceleration deflects the horizontal path of any object moving on the Earth's surface. As a result it affects the directions of wind and ocean currents, eventually constraining them to form circulatory patterns about centers of high or low pressure, and thereby plays an important role in determining the weather. In the Earth's core it has an important influence on the fluid motions that generate the geomagnetic field.

3.3.4 The Tides

The gravitational forces of the Sun and Moon deform the Earth's shape, causing tides in the oceans, atmosphere, and solid body of the Earth. The most visible tidal effects are the displacements of the ocean surface, which is a hydrostatic equipotential surface. The Earth does not react rigidly to the tidal forces. The solid body of the Earth deforms in a like manner to the free surface, giving rise to *bodily Earth tides*. These can be observed with specially designed instruments, which operate on a similar principle to the long-period seismometer (Section 6.5.3.2).

The height of the marine equilibrium tide amounts to only half a meter or so over the free ocean. In coastal areas the tidal height is significantly increased by the shallowing of the continental shelf and the confining shapes of bays and harbors. Accordingly, the height and variation of the tide at any place are influenced strongly by complex local factors. Subsequent chapters deal with the tidal deformations of the Earth's hydrostatic figure.

3.3.4.1 Lunar Tidal Periodicity

The Earth and Moon are coupled together by gravitational attraction. Their common motion is like that of a pair of ballroom dancers. Each partner moves around the center of mass (or *barycenter*) of the pair. For the Earth–Moon pair the location of the center of mass is easily found. Let E be the mass of the Earth and m that of the Moon; let the separation of the centers of the Earth and Moon be r_L and let the distance of their common center of mass be d from the center of the Earth. If their common rate of rotation is Ω, the centrifugal force of the Earth about the center of mass is $Ed\Omega^2$ and that of the Moon is $m(r_L - d)\Omega^2$. Equating these forces we get

$$d = \frac{m}{E+m} r_L.$$

(3.34)

The mass of the Moon is 0.0123 that of the Earth and the distance between the centers is 384,100 km. These figures give $d = 4600$ km, i.e., the center of revolution of the Earth–Moon pair lies within the Earth.

It follows that the paths of the Earth and the Moon around the sun are more complicated than at first appears. The elliptical orbit is traced out by the *barycenter* of the pair (Fig. 3.10). The Earth and Moon follow wobbly paths, which, while always concave toward the Sun, bring each body at different times of the month alternately inside and outside the elliptical orbit.

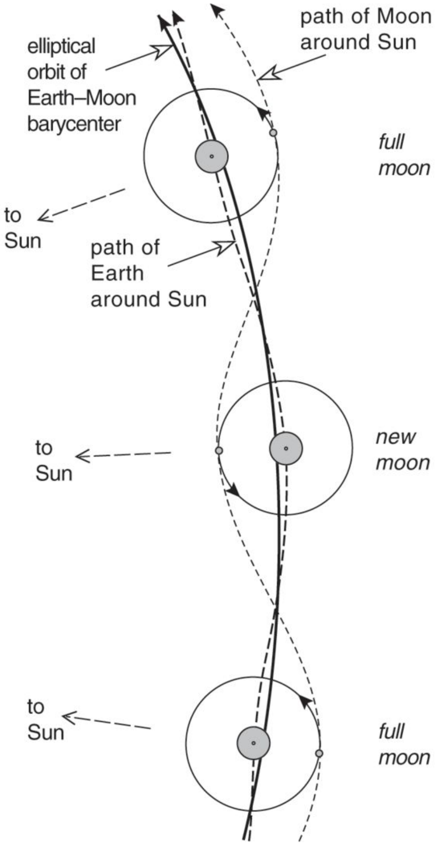

elliptical orbit of Earth–Moon barycenter
path of Moon around Sun
full moon
to Sun
path of Earth around Sun
to Sun
new moon
to Sun
full moon

Fig. 3.10 Paths of the Earth and Moon, and their barycenter, around the Sun.

To understand the common revolution of the Earth–Moon pair we have to exclude the rotation of the Earth about its axis. The "revolution without rotation" is illustrated in Fig. 3.11. The Earth–Moon pair revolves about S, the center of mass. Let the starting positions be as shown in Fig. 3.11a. Approximately one week later the Moon has advanced in its path by one-quarter of a revolution and the center of the Earth has moved so as to keep the center of mass fixed (Fig. 3.11b). The relationship is maintained in the following weeks (Fig. 3.11c, d) so that during one month the center of the Earth describes a circle about S. Now consider the motion of point number 2 on the left-hand side of the Earth in Fig. 3.11. If the Earth revolves as a rigid body and the rotation about its own axis is omitted, after one week point 2 will have moved to a new position but will still be the furthest point on the left. Subsequently, during one month point 2 will describe a small circle with the same radius as the circle described by the Earth's center. Similarly points 1, 3, and 4 will also describe circles of exactly the same size. A simple illustration of this point can be made by chalking the tip of each finger on one hand with a different color, then moving your hand in a circular motion while touching a blackboard; your fingers will draw a set of identical circles.

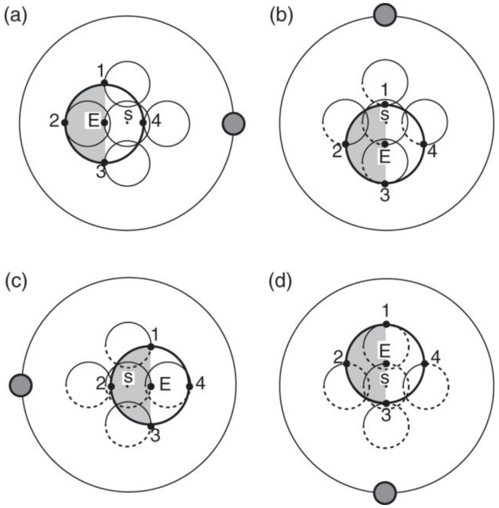


Fig. 3.11 Illustration of the "revolution without rotation" of the Earth–Moon pair about their common center of mass at S.

The “revolution without rotation” causes each point in the body of the Earth to describe a circular path with identical radius. The centrifugal acceleration of this motion has therefore the same magnitude at all points in the Earth, and, as can be seen by inspection of Fig. 3.11a–d, it is directed away from the Moon and parallel to the Earth–Moon line of centers. At the center of the Earth (C in Fig. 3.12a) this centrifugal acceleration exactly balances the gravitational attraction of the Moon. Its magnitude is given by

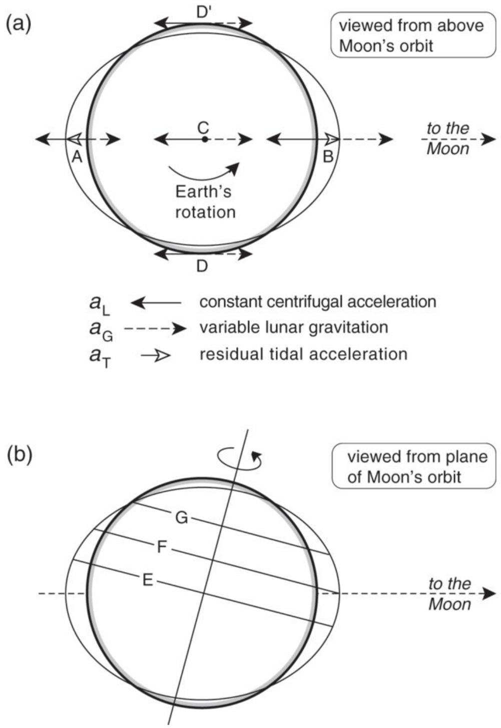

(a)
D'
viewed from above Moon's orbit
C
A
B
to the Moon
Earth's rotation
D
a_L constant centrifugal acceleration
a_G variable lunar gravitation
a_T residual tidal acceleration
(b)
viewed from plane of Moon's orbit
G
F
E
to the Moon

Fig. 3.12 (a) The relationships of the centrifugal, gravitational, and residual tidal accelerations at selected points in the Earth. (b) Latitude effect that causes diurnal inequality of the tidal height.

$$a_{\mathrm{L}} = G\frac{m}{r_{\mathrm{L}}^2}. \tag{3.35}$$

At B, on the side of the Earth nearest to the Moon, the gravitational acceleration of the Moon is larger than at the center of the Earth and exceeds the centrifugal acceleration a_{L}. There is a residual acceleration *toward* the Moon, which raises a tide on this side of the Earth. The magnitude of the tidal acceleration at B is

$$a_{\mathrm{T}} = Gm\left(\frac{1}{(r_{\mathrm{L}} - R)^2} - \frac{1}{r_{\mathrm{L}}^2}\right),$$
(3.36)

$$a_{\mathrm{T}} = \frac{Gm}{r_{\mathrm{L}}^2}\left(\left(1 - \frac{R}{r_{\mathrm{L}}}\right)^{-2} - 1\right).$$
(3.37)

Expanding this equation with the binomial theorem and simplifying gives

$$a_{\mathrm{T}} = \frac{Gm}{r_{\mathrm{L}}^2}\left(2\frac{R}{r_{\mathrm{L}}} + 3\left(\frac{R}{r_{\mathrm{L}}}\right)^2 + \cdots\right).$$
(3.38)

At A, on the far side of the Earth, the gravitational acceleration of the Moon is less than the centrifugal acceleration a_{L}. The residual acceleration (Fig. 3.12a) is *away from* the Moon, and raises a tide on the far side of the Earth. The magnitude of the tidal acceleration at A is

$$a_{\mathrm{T}} = Gm\left(\frac{1}{r_{\mathrm{L}}^2} - \frac{1}{(r_{\mathrm{L}} + R)^2}\right),$$
(3.39)

which can be approximated by

$$a_T = \frac{Gm}{r_L^2}\left(2\frac{R}{r_L} - 3\left(\frac{R}{r_L}\right)^2 + \cdots\right). \tag{3.40}$$

At points D and D′ the direction of the gravitational acceleration due to the Moon is not exactly parallel to the line of centers of the Earth–Moon pair. The residual tidal acceleration is almost along the direction toward the center of the Earth. Its effect is to lower the free surface in this direction.

The free hydrostatic surface of the Earth is an equipotential surface (Section 3.2.3), which in the absence of the Earth's rotation, tidal effects, and lateral density variations would be a sphere. The lunar tidal accelerations perturb the equipotential surface, raising it at A and B while lowering it at D and D′, as in Fig. 3.12a. The tidal deformation of the Earth produced by the Moon thus has an almost prolate ellipsoidal shape, like a rugby ball, along the Earth–Moon line of centers. The daily tides are caused by superposing the Earth's rotation on this deformation. In the course of one day a point rotates past points A, D, B, and D′ and an observer experiences two full tidal cycles, called the *semi-diurnal* tides. The extreme tides are not equal at every latitude, because of the varying angle between the Earth's rotational axis and the Moon's orbit (Fig. 3.12b). At the equator E the semi-diurnal tides are equal; at an intermediate latitude F one tide is higher than the other; and at latitude G and higher there is only one (diurnal) tide per day. The difference in height between two successive high or low tides is called the *diurnal inequality*.

In the same way that the Moon deforms the Earth, so the Earth causes a tidal deformation of the Moon. In fact, the tidal relationship between any planet and one of its moons, or between the Sun and a planet or comet, can be treated analogously to the Earth–Moon pair. A tidal acceleration similar to Eq. (3.38) deforms the smaller body; its self-gravitation acts to counteract the deformation. However, if a moon or comet comes too close to the planet, the tidal forces deforming it may overwhelm the gravitational forces holding it together, so that the moon or comet is torn apart. The separation at which this occurs is the *Roche limit* (Box 3.2) of the planet. The material of a disintegrated moon or comet enters orbit around the planet, forming a system of concentric rings, as around the giant planets (Section 1.6).

Box 3.2 The Roche Limit

Suppose that a moon with mass M and radius R_M is in orbit at a distance d from a planet with mass P and radius R_P. The Roche limit is the distance at which the tidal attraction exerted by the planet on the moon overcomes the moon’s self-gravitation (Fig. B3.2.1). If the moon is treated as an elastic body, its deformation to an elongate form complicates the calculation of the Roche limit. However, for a rigid body, the computation is simple because the moon maintains its shape as it approaches the planet.

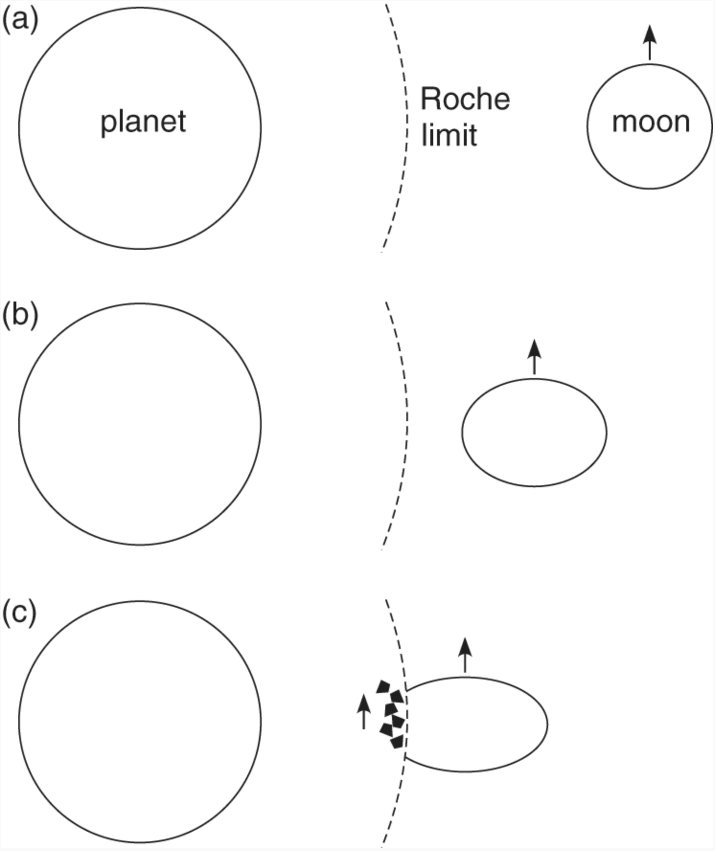


Fig. B3.2.1 (a) Far from its parent planet, a moon is spherical in shape, but (b) as it comes closer tidal forces deform it into an ellipsoidal shape, until (c) within the Roche limit the moon breaks up. The disrupted material forms a ring of small objects orbiting the planet in the same sense as the moon's orbital motion.

Consider the forces acting on a small mass m forming part of the rigid moon's surface closest to the planet (Fig. B3.2.2). The tidal acceleration a_T caused by the planet can be written by adapting the first term of Eq. (3.38), which gives the deforming force F_T on the small mass,

$$F_T = ma_T = G\frac{mP}{d^2}\left(2\frac{R_M}{d}\right) = 2G\frac{mPR_M}{d^3}. \tag{1}$$

This disrupting force is counteracted by the gravitational force F_{G} of the moon, which is

$$F_{\mathrm{G}} = ma_{\mathrm{G}} = G\frac{mM}{(R_{\mathrm{M}})^2}.$$

(2)

The Roche limit d_{R} for a rigid solid body is determined by equating these forces:

$$2G\frac{mPR_{\mathrm{M}}}{(d_{\mathrm{R}})^3} = G\frac{mM}{(R_{\mathrm{M}})^2},$$

(3)

$$(d_R)^3 = 2\frac{P}{M}(R_M)^3.$$

(4)

If the densities of the planet, ρ_{P}, and moon, ρ_{M}, are known, Eq. (4) can be rewritten:

$$(d_{\mathrm{R}})^3 = 2\frac{\left(\frac{4}{3}\pi\rho_{\mathrm{P}}(R_{\mathrm{P}})^3\right)}{\left(\frac{4}{3}\pi\rho_{\mathrm{M}}(R_{\mathrm{M}})^3\right)}(R_{\mathrm{M}})^3 = 2\left(\frac{\rho_{\mathrm{P}}}{\rho_{\mathrm{M}}}\right)(R_{\mathrm{P}})^3,$$

(5)

$$d_{\mathrm{R}} = R_{\mathrm{P}}\left(2\frac{\rho_{\mathrm{P}}}{\rho_{\mathrm{M}}}\right)^{1/3} = 1.26R_{\mathrm{P}}\left(\frac{\rho_{\mathrm{P}}}{\rho_{\mathrm{M}}}\right)^{1/3}.$$

(6)

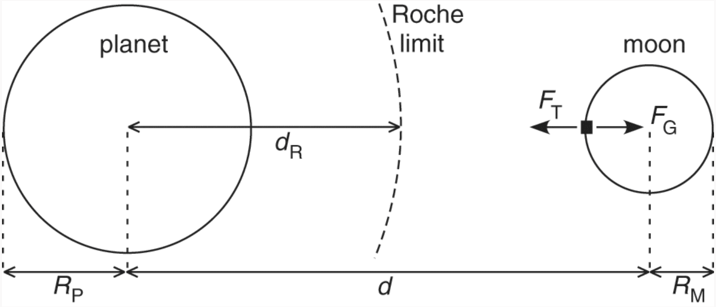


Fig. B3.2.2 Parameters for computation of the Roche limit.

If the moon is fluid, tidal attraction causes it to elongate progressively as it approaches the planet. This complicates the exact calculation of the Roche limit, but it is given approximately by

$$d_R == 2.42 R_P \left(\frac{\rho_P}{\rho_M} \right)^{1/3} . \quad (7)$$

Comparison of Eq. (6) and Eq. (7) shows that a fluid or gaseous moon disintegrates about twice as far from the planet as a rigid moon. In practice, the Roche limit for a moon about its parent planet (and the planet about the Sun) depends on the rigidity of the moon and lies between the two extremes.

3.3.4.2 Tidal Effect of the Sun

The Sun also has an influence on the tides. The theory of the solar tides can be followed in an identical manner to the lunar tides by again applying the principle of "revolution without rotation." The Sun's mass is 333,000 times greater than that of the Earth, so the common center of mass is close to the center of the Sun at a radial distance of about 450 km from its center. The period of the revolution is one year. As for the lunar tide, the imbalance between gravitational acceleration of the Sun and centrifugal acceleration due to the common revolution leads to a prolate ellipsoidal tidal deformation. The solar effect is smaller than that of the Moon. Although the mass of the Sun is vastly greater than that of the Moon, its distance from the Earth is also much greater. Equation (3.38) shows that, to first order, the effective tidal acceleration varies inversely with the *cube* of distance, and consequently the maximum tidal effect of the Sun is only about 45% that of the Moon.

3.3.4.3 Spring and Neap Tides

The superposition of the lunar and solar tides causes a modulation of the tidal amplitude. The ecliptic plane is defined by the Earth's orbit around the Sun. The Moon's orbit around the Earth is not exactly in the ecliptic but is inclined at a very small angle of about 5° to it. For discussion of the combination of lunar and solar tides we can assume the orbits to be coplanar. The Moon and Sun each produce a prolate tidal deformation of the Earth, but the relative orientations of these ellipsoids vary during one month (Fig. 3.13). At conjunction the (new) Moon is on the same side of the Earth as the Sun, and the ellipsoidal deformations augment each other. This is also the case half a month later at opposition, when the (full) Moon is on the opposite side of the Earth from the Sun. The unusually high tides at opposition and conjunction are called *spring tides*. In contrast, at the times of quadrature the waxing or waning half Moon causes a prolate ellipsoidal deformation out of phase with the solar deformation. The maximum lunar tide coincides with the minimum solar tide, and the effects partially cancel each other. The unusually low tides at quadrature are called *neap tides*. The superposition of the lunar and solar tides causes modulation of the tidal amplitude during a month (Fig. 3.14).

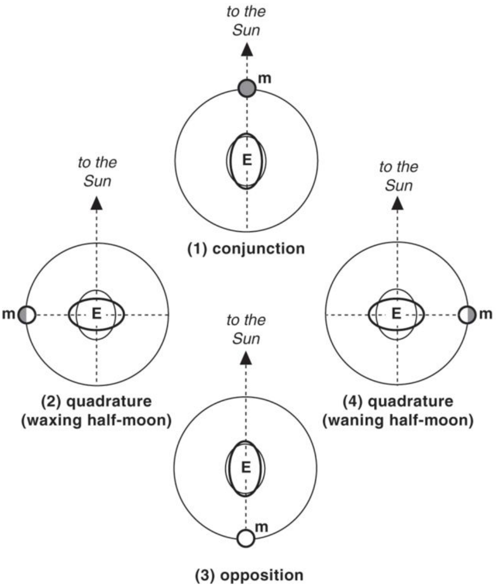

to the Sun
m
E
(1) conjunction
to the Sun
m
E
(2) quadrature
(waxing half-moon)
to the Sun
E
m
(4) quadrature
(waning half-moon)
to the Sun
E
m
(3) opposition

Fig. 3.13 The orientations of the solar and lunar tidal deformations of the Earth at different lunar phases.

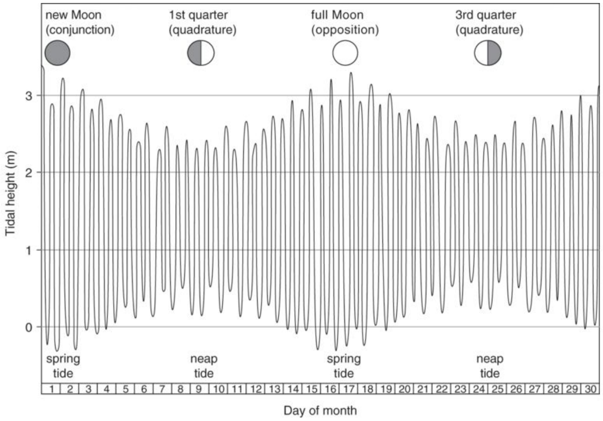


Fig. 3.14 Schematic representation of the modulation of the tidal amplitude as a result of superposition of the lunar and solar tides.

3.3.4.4 Effect of the Tides on Gravity Measurements

The tides have an effect on gravity measurements made on the Earth. The combined effects of the Sun and Moon cause an acceleration at the Earth's surface of approximately 0.3 mGal, of which about two-thirds are due to the Moon and one-third to the Sun. The sensitive modern instruments used for gravity exploration can readily detect gravity differences of 0.01 mGal. It is necessary to compensate gravity measurements for the tidal effects, which vary with location, date and time of day. Fortunately, tidal theory is so well established that the gravity effect can be calculated and tabulated for any place and time before beginning a survey.

3.3.4.5 Bodily Earth Tides

A simple way to measure the height of the marine tide might be to fix a stake to the sea bottom at a suitably sheltered location and to record continuously the measured water level (assuming that confusion introduced by wave motion can be eliminated or taken into account). The observed amplitude of the marine tide, defined by the displacement of the free water surface, is found to be about 70% of the theoretical value. The difference is explained by the elasticity of the Earth. The tidal deformation corresponds to a redistribution of mass, which modifies the gravitational potential of the Earth and augments the elevation of the free surface. This is partially counteracted by a bodily tide in the solid Earth, which deforms elastically in response to the attraction of the Sun and Moon. The free water surface is raised by the tidal attraction, but the sea bottom in which the measuring rod is implanted is also raised. The measured tide is the difference between the marine tide and the bodily Earth tide.

The displacement of the equipotential surface can be measured with a horizontal pendulum, which reacts to the tilt of the surface. The bodily Earth tides affect gravity measurements and thus can be observed with sensitive gravimeters. Their effects are incorporated into the tidal corrections that are applied to gravity measurements.

3.3.4.6 Love's Numbers

In 1911 A. E. H. Love introduced a set of numbers to describe numerically the response of the Earth to the Moon's tidal forces. The Love numbers can be adapted to any planet as a way to express its rigidity and its deformation due to tidal forces.

The free equipotential surface of the Earth is modified by a tidal potential, which has a prolate shape, reflecting the shape of the Moon's deforming potential (Fig. 3.12). Love showed that the extra potential of the tidal deformation is proportional to the potential of the deforming forces, with a constant of proportionality denoted k_2 and referred to as a Love number. The subscript 2 indicates that the number is calculated from the second-degree spherical harmonic function (Box 3.4) that describes the deforming potential. If the height of the equilibrium tide, without taking elasticity into account, is H, the elasticity raises the tidal height by the amount k_2H (Fig. 3.15). The value of k_2 is a measure of the elastic response of the Earth as a whole. For a rigid planet, $k_2 = 0$ because it does not deform, while for a homogeneous fluid body $k_2 = 1.5$; the measured value for the Earth is $k_2 = 0.298$.

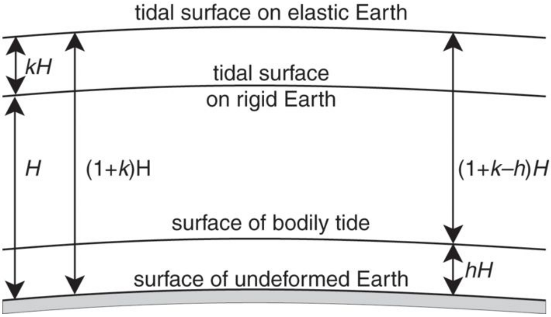


Fig. 3.15 The role of the Love numbers k and h in evaluating the heights of the marine and bodily tides.

In a similar way, Love defined a second number, h_2, which relates the height of the solid Earth tide to the lunar deforming potential and to the equilibrium tide, so that the height of the bodily tide is h_2H. The Love number h_2 is therefore an expression of the global elasticity of the solid Earth. For a rigid planet $h_2 = 0$ and for a fluid body $h_2 = 1$; the measured value for the Earth is $h_2 = 0.603$.

The Love numbers h_2 and k_2 are related by a linear equation to the height of the marine tides; they are also related by a different linear equation to the effect of the tides on gravity. The values of k_2 and h_2 are obtained by solving the two equations simultaneously. The magnitude of the tidal effect on gravity can thus be calculated for any location, which allows tidal corrections to be made to sensitive gravity measurements.

The vertical displacement of the equipotential surface beneath the position of the Moon requires a compensating horizontal displacement. The horizontal tide, although much weaker than the vertical tide, is also proportional to the lunar deforming potential and is expressed by the number l, which has the value 0.084 for the Earth. Although l was introduced by the Japanese scientist T. Shida in 1912, it is often grouped with h and k as one of the Love numbers.

The Love numbers play an important role in understanding the yielding of a planet to deforming forces, such as those responsible for the tides or which result from the planet's rotation. For example, the yielding of the Earth to rotational forces lengthens the period of the Chandler wobble (Section 3.3.5.3) by about 40%. Comparison of the observed period with that for a rigid body yields a value $k_2 \approx 0.28$ for the Earth, close to the value from tidal observations. In planetary investigations the magnitudes of the Love numbers, especially of k_2, are indicators of the elasticity of a planet as a whole and of its internal mass distribution.

A spherical harmonic model to degree and order 40 of the gravity field of Mercury was obtained from years of Earth-based radio tracking of the MESSENGER spacecraft. The annual solar tide due to Mercury's orbit around the Sun deforms the planet's shape and modulates its gravitational field, from which the Love number k_2 can be calculated. This gave a value $k_2 = 0.464$ for Mercury, which is compatible with an internal structure of an outer solid shell overlying a thin liquid outer core, which decouples the shell from a solid inner core.

3.3.5 Changes in the Earth's Rotation

The Earth's rotation is affected by the gravitational attractions of the Sun, Moon, and planets. The rate of rotation and the orientation of the rotational axis change with time. The orbital motion around the Sun is also affected. The orbit rotates about the pole to the plane of the ecliptic and its ellipticity changes over long periods of time.

3.3.5.1 Effect of Lunar Tidal Friction on the Length of the Day

If the Earth reacted perfectly elastically to the lunar tidal forces, the prolate tidal bulge would be aligned along the line of centers of the Earth–Moon pair (Fig. 3.16a). However, the motion of the seas is not instantaneous and the tidal response of the solid part of the Earth is partly anelastic (Section 5.3.1). These features cause a slight delay in the time when high tide is reached, amounting to about 12 minutes. In this short interval the Earth's rotation carries the line of the maximum tides past the line of centers by a small angle of approximately 2.9° (Fig. 3.16b). A point on the rotating Earth passes under the line of maximum tides 12 minutes after it passes under the Moon. The small phase difference is called the tidal lag.

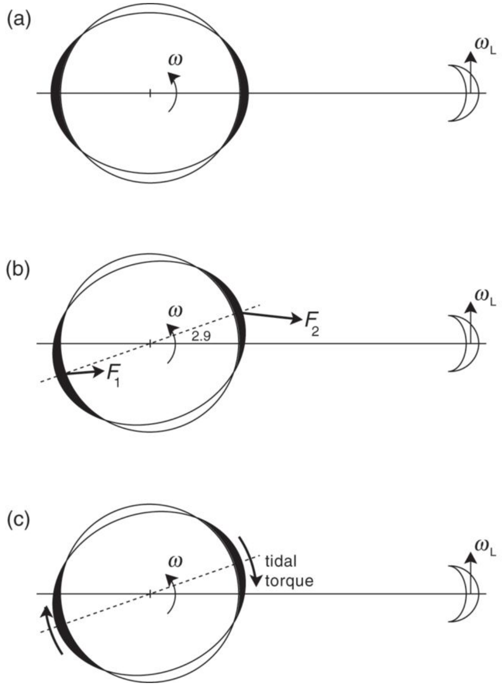


Fig. 3.16 (a) Alignment of the prolate tidal bulge of a perfectly elastic Earth along the line of centers of the Earth–Moon pair. (b) Tidal phase lag of 2.9° relative to the line of centers due to the Earth's partially anelastic response. (c) Tidal decelerating torque due to unequal gravitational attractions of the Moon on the far- and near-sided tidal bulges.

Because of the tidal lag the gravitational attraction of the Moon on the tidal bulges on the far side and near side of the Earth (F_1 and F_2, respectively) are not collinear (Fig. 3.16b). F_2 is stronger than F_1 so a torque is produced in the opposite sense to the Earth's rotation (Fig. 3.16c). The tidal torque acts as a brake on the Earth's rate of rotation, which is gradually slowing down.

The tidal deceleration of the Earth is manifested in a gradual increase in the length of the day. The effect is very small. Tidal theory predicts an increase in the length of the day of only 2.4 ms per century. Observations of the phenomenon are based on ancient historical records of lunar and solar eclipses and on telescopically observed occultations of stars by the Moon. The current rate of rotation of the Earth can be measured with very accurate atomic clocks. Telescopic observations of the daily times of passage of stars past the local zenith are recorded with a camera controlled by an atomic clock. These observations give precise measures of the mean value and fluctuations of the length of the day.

The occurrence of a lunar or solar eclipse was a momentous event for ancient peoples, and was duly recorded in scientific and non-scientific chronicles. Untimed observations are found in non-astronomical works. They record, with variable reliability, the degree of totality and the time and place of observation. The unaided human eye is able to decide quite precisely just when an eclipse becomes total. Timed observations of both lunar and solar eclipses made by Arab astronomers around AD 800–1000 and Babylonian astronomers a thousand years earlier give two important groups of data (Fig. 3.17). By comparing the observed times of alignment of the Sun, Moon, and Earth with times predicted from the theory of celestial mechanics, the differences due to change in length of the day may be computed. A straight line with slope equal to the rate of increase of the length of the day inferred from tidal theory, 2.4 ms per century, connects the Babylonian and Arab datasets. Since the medieval observations of Arab astronomers, the length of the day has increased on average by about 1.4 ms per century. The dataset based on telescopic observations covers the time from AD 1620 to 1980. It gives a more detailed picture and shows that the length of the day fluctuates about the long-term trend of 1.4 ms per century. A possible interpretation of the difference between the two slopes is that non-tidal causes have opposed the deceleration of the Earth's rotation since about AD 950. It would be wrong to infer that some sudden event at that epoch caused an abrupt change, because the data are equally compatible with a smoothly changing polynomial. The observations confirm the importance of tidal braking, but they also indicate that tidal friction is not the only mechanism affecting the Earth's rotation.

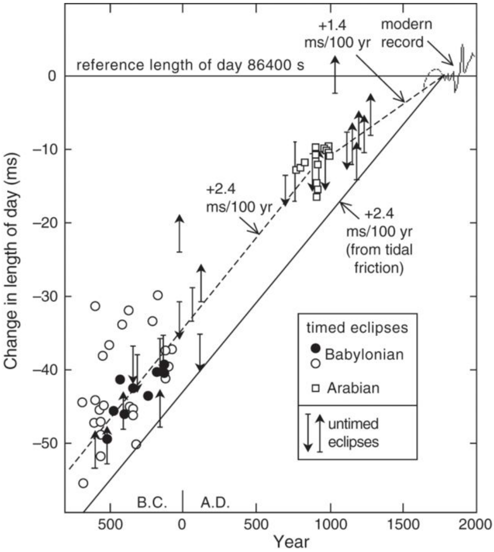


Fig. 3.17 Long-term changes in the length of the day deduced from observations of solar and lunar eclipses between 700 BC and AD 1980.

Source: F.R. Stephenson and L.V. Morrison, Long-term changes in the rotation of the earth: 700 B.C. to A.D. 1980, *Phil. Trans R. Soc. Lond. A* 313, 47–70, 1984.

The short-term fluctuations in rotation rate are due to exchanges of angular momentum with the Earth’s atmosphere and core. The atmosphere is tightly coupled to the solid Earth. An increase in average global wind speed corresponds to an increase in the angular momentum of the atmosphere and corresponding decrease in angular momentum of the solid Earth. Accurate observations by very long baseline interferometry (see Section 3.5.8) confirm that rapid fluctuations in the length of the day are directly related to changes in the angular momentum of the atmosphere.

In the course of seconds or minutes, the rotation rate of the Earth may change measurably in response to large earthquakes. The redistribution of mass caused by earthquakes modifies the Earth’s moments of inertia. To ensure that the angular momentum remains constant, this is compensated by a change in rotation rate and the instantaneous position of the rotation axis.

On a longer timescale of decades, the changes in length of the day may be related to changes in the angular momentum of the core. The fluid in the outer core has a speed of the order of 0.1 mm s^{-1} relative to the overlying mantle. The mechanism for exchange of angular momentum between the fluid core and the rest of the Earth depends on the way the core and mantle are coupled. The coupling may be mechanical if topographic irregularities obstruct the flow of the core fluid along the core–mantle interface. The core fluid is a good electrical conductor, so if the lower mantle also has an appreciable electrical conductivity it is possible that the core and mantle are coupled electromagnetically.

3.3.5.2 Increase of the Earth–Moon Distance

Further consequences of lunar tidal friction can be seen by applying the law of conservation of momentum to the Earth–Moon pair. Let the Earth's mass be E, its rate of rotation be ω, and its moment of inertia about the rotation axis be C; let the corresponding parameters for the Moon be m, Ω_L, and C_L, and let the Earth–Moon distance be r_L. Further, let the distance of the common center of revolution from the center of the Earth be d, as given by Eq. (3.34). The angular momentum of the system is given by

$$C\omega + E\Omega_L d^2 + m\Omega_L (r_L - d)^2 + C_L\Omega_L = \text{constant.} \tag{3.41}$$

The fourth term is the angular momentum of the Moon about its own axis. Tidal deceleration due to the Earth's attraction has slowed down the Moon's rotation until it equals its rate of revolution about the Earth. Both Ω_L, and C_L are very small and the fourth term can be neglected. The second and third terms can be combined so that we get

$$C\omega + \left(\frac{E}{E+m}\right) m\Omega_L r_L^2 = \text{constant.}$$

(3.42)

The gravitational attraction of the Earth on the Moon is equal to the centripetal acceleration of the Moon about the common center of revolution, thus

$$G\frac{E}{r_{\mathrm{L}}^{2}}=\Omega_{\mathrm{L}}^{2}(r_{\mathrm{L}}-d)=\Omega_{\mathrm{L}}^{2}r_{\mathrm{L}}\left(\frac{E}{E+m}\right),\tag{3.43}$$

from which

$$\Omega_{\mathrm{L}}r_{\mathrm{L}}^{2}=\sqrt{G(E+m)r_{\mathrm{L}}}.\tag{3.44}$$

Inserting this in Eq. (3.42) gives

$$C\omega+\frac{Em}{\sqrt{(E+m)}}\sqrt{Gr_{\mathrm{L}}}=\text{constant}.\tag{3.45}$$

The first term in this equation decreases, because tidal friction reduces ω. To conserve angular momentum the second term must increase. Thus, lunar tidal braking of the Earth's rotation causes an increase in the Earth–Moon distance, r_{L}. At present this distance is increasing at about 38 mm yr^{-1}. As a further consequence, Eq. (3.44) shows that the Moon's rate of revolution about the Earth (Ω_{L}) – and consequently also its synchronous rotation about its own axis – must decrease when r_{L} increases. Thus, tidal friction slows down the rates of Earth rotation, lunar rotation, and lunar orbital revolution, and increases the Earth–Moon distance.

Eventually a situation will evolve in which the Earth's rotation has slowed until it is synchronous with the Moon's own rotation and its orbital revolution about the Earth. All three rotations will then be synchronous and equivalent to about 48 present Earth days. This will happen when the Moon's distance from the Earth is about 88 times the Earth's radius ($r_L = 88\,R$; it is presently equal to about 60 R). The Moon will then be stationary over the Earth, and Earth and Moon will constantly present the same face to each other. This configuration already exists between the planet Pluto and its satellite Charon.

3.3.5.3 The Chandler Wobble

The Earth's rotation gives it the shape of a spheroid, or ellipsoid of revolution. This figure is symmetric with respect to the mean axis of rotation, about which the moment of inertia is greatest; this is also called the axis of figure (see Section 3.4). However, at any moment the instantaneous rotational axis is displaced by a few meters from the axis of figure. The orientation of the total angular momentum vector remains nearly constant but the axis of figure changes location with time and meanders around the rotation axis: the chart of the polar motion is called a *polhody* (Fig. 3.18).

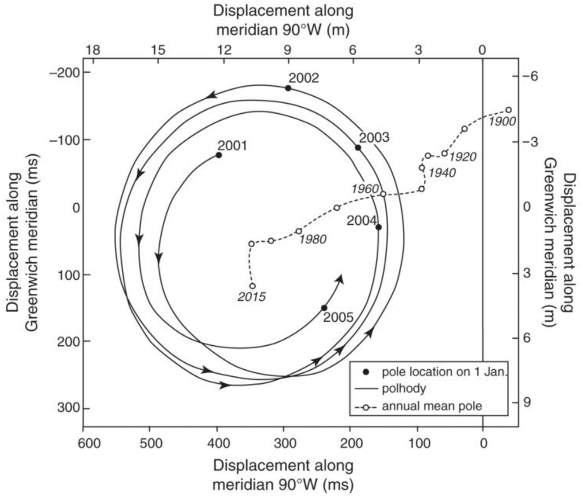


Fig. 3.18 Chart of the motion of Earth's instantaneous rotation pole from 2001 to 2005 and the long-term drift of the mean pole position from 1900 to 2015. The irregular quasi-circular motion is due to superposition of the free Chandler wobble with a period of 435 days and a forced annual component resulting from seasonal

displacements in the atmosphere and oceans.

Data source: International Earth Rotation and Reference Systems Service.

The theory of this motion was described by Leonhard Euler (1707–1783), a Swiss mathematician. He showed that the displaced rotational axis of a rigid spheroid would execute a circular motion about its mean position, now called the *Euler nutation*. Because it occurs in the absence of an external driving torque, it is also called the *free nutation*. It is due to differences in the way mass is distributed about the axis of rotational symmetry and an axis at right angles to it in the equatorial plane. The mass distributions are represented by the moments of inertia about these axes. If C and A are the moments of inertia about the rotational axis and an axis in the equatorial plane, respectively, Euler's theory shows that the period of free nutation is $A/(C-A)$ days, or approximately 305 days.

Astronomers were initially unsuccessful in detecting a polar motion with this period. However, in 1891 an American geodesist and astronomer, S. C. Chandler, reported that the polar motion of the Earth's axis contained two important components. An annual component with amplitude about 0.10 seconds of arc is forced by the transfer of mass between atmosphere and hydrosphere accompanying the changing of the seasons. A slightly larger component with amplitude 0.15 seconds of arc has a period of 435 days. This polar motion is now called the *Chandler wobble*. It corresponds to the free Euler nutation in an elastic Earth. The increase in period from 305 days to 435 days is a consequence of the elastic yielding of the Earth. The superposition of the annual and Chandler frequencies results in a "beat effect," in which the amplitude of the latitude variation is modulated with a period of 6–7 years.

The Chandler wobble is routinely observed by means of very precise methods of space geodesy (Section 3.5), such as very long baseline interferometry (VLBI), laser ranging, and the Global

Navigation Satellite System (GNSS).

3.3.5.4 Precession and Nutation of the Rotation Axis

During its orbital motion around the Sun, the Earth's axis maintains an (almost) constant tilt of about 23.5° to the pole to the ecliptic. The line of intersection of the plane of the ecliptic with the equatorial plane is called the line of equinoxes. Two times a year, when this line points directly at the Sun, day and night have equal duration over the entire globe.

In the theory of the tides the unequal lunar attractions on the near- and far-sided tidal bulges cause a torque about the rotation axis, which has a braking effect on the Earth's rotation. The attractions of the Moon (and Sun) on the equatorial bulge due to rotational flattening also produce torques on the spinning Earth. On the side of the Earth nearer to the Moon (or Sun) the gravitational attraction F_2 on the equatorial bulge is greater than the force F_1 on the distant side (Fig. 3.19a). Due to the tilt of the rotation axis to the ecliptic plane, the forces are not collinear. A torque results, which acts about a line in the equatorial plane, normal to the Earth–Sun line and normal to the spin axis. The magnitude of the torque changes as the Earth orbits around the Sun. It is minimum (and zero) at the spring and autumn equinoxes and maximum at the summer and winter solstices.

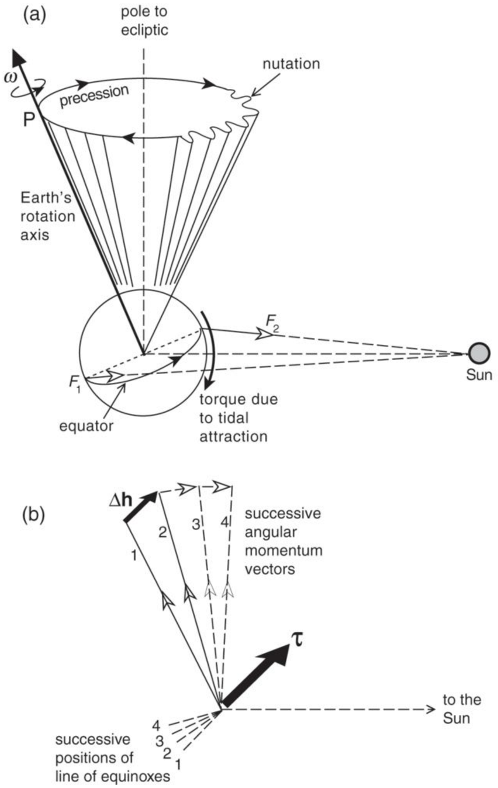

(a)
pole to ecliptic
nutation
ω
precession
P
Earth's rotation axis
F_2
Sun
F_1
equator
torque due to tidal attraction
(b)
Δh
1
2
3
4
successive angular momentum vectors
τ
to the Sun
successive positions of line of equinoxes
4
3
2
1

Fig. 3.19 (a) The precession and forced nutation (greatly exaggerated) of the rotation axis due to the lunar torque on the spinning Earth (after Strahler, 1963). (b) Torque and incremental angular momentum changes resulting in precession.

The response of a rotating system to an applied torque is to acquire an additional component of angular momentum parallel to the torque. In our example this will be perpendicular to the angular momentum (h) of the spinning Earth. The torque has a component (τ) parallel to the line of equinoxes (Fig. 3.19b) and a component normal to this line in the equatorial plane. The torque τ causes an increase Δh in angular momentum and shifts the angular momentum vector to a new position. If this exercise is repeated incrementally, the rotation axis moves around the surface of a cone whose axis is the pole to the ecliptic (Fig. 3.19a). The geographic pole P moves around a circle in the opposite sense from the Earth's spin. This motion is called *retrograde precession*. It is not a steady motion, but pulsates in sympathy with the driving torque. A change in orientation of the rotation axis affects the location of the line of equinoxes and causes the timing of the equinoxes to change slowly. The rate of change is only 50.4 arcsec per year, but it has been recognized during centuries of observation. For example, the Earth's rotation axis now points at Polaris in the constellation Ursa Minor, but in the time of the Egyptians around 3000 BC the pole star was Alpha Draconis, the brightest star in the constellation Draco. Hipparchus is credited with discovering the precession of the equinoxes in 120 BC by comparing his own observations with those of earlier astronomers.

The theory of the phenomenon is well understood. The Moon also exerts a torque on the spinning Earth and contributes to the precession of the rotation axis (and equinoxes). As in the theory of the tides, the small size of the Moon compared to the Sun is more than compensated by its nearness, so that the precessional contribution of the Moon is about double the effect of the Sun. The theory of precession shows that the period of 25,700 years is proportional to the Earth's *dynamical ellipticity*, *H* (Section 3.4.1). This ratio (equal to 1/305.4413) is an important indicator of the internal distribution of mass in the Earth.

The component of the torque in the equatorial plane adds an additional motion to the axis, called *nutation*, because it causes the axis to nod up and down (Fig. 3.18a). The solar torque causes a semi-annual nutation, the lunar torque a semi-monthly one. In fact, the motion of the axis exhibits many *forced nutations*, so-called because they respond to external torques. All are tiny perturbations on the precessional motion, the largest having an amplitude of only about 9 arcsec and a period of 18.6 years. This nutation results from the fact that the plane of the lunar orbit is inclined at 5.145° to the plane of the ecliptic and (like the motion of artificial Earth satellites) precesses retrogradely. This causes the inclination of the lunar orbit to the equatorial plane to vary between about 18.4° and 28.6°, modulating the torque and forcing a nutation with a period of 18.6 years.

It is important to note that the Euler nutation and Chandler wobble are polar motions about the rotation axis, but the precession and forced nutations are displacements of the rotation axis itself.

3.3.5.5 Milankovitch Climatic Cycles

Solar energy can be imagined as flowing equally from the Sun in all directions. At distance r it floods a sphere with surface area $4\pi r^2$. The amount of solar energy falling per second on a square meter (the *insolation*) therefore decreases as the inverse square of the distance from the Sun. The gravitational attractions of the Moon, Sun, and the other planets – especially Jupiter – cause cyclical changes of the orientation of the rotation axis and variations in the shape and orientation of Earth's orbit. These variations modify the insolation of the Earth and result in long-term periodic changes in Earth's climate.

The angle between the rotational axis and the pole to the ecliptic is called the *obliquity*. It is the main factor determining the seasonal difference between summer and winter in each hemisphere. In the northern hemisphere, the insolation is maximum at the summer solstice (currently June 21) and minimum at the winter solstice (December 21–22). The exact dates change with the precession of the equinoxes, and also depend on the occurrence of leap years. The solstices do not coincide with extreme positions in the Earth's orbit. The Earth currently reaches aphelion, its furthest distance from the Sun, around July 4–6, shortly after the summer solstice, and passes perihelion around January 2–4. About 13,000 years from now, as a result of precession, the summer solstice will occur when Earth is close to perihelion. In this way, precession causes long-term changes in climate with a period related to the precession.

The gravitational attraction of the other planets causes the obliquity to change cyclically with time. It is currently equal to $23^\circ\,26'\,21.4''$ but varies slowly between a minimum of $21^\circ\,55'$ and a maximum of $24^\circ\,18'$. When the obliquity increases, the seasonal differences in temperature become more pronounced, while the opposite effect ensues if obliquity decreases. Thus, the variation in obliquity causes a modulation in the seasonal contrast between summer and winter on a global scale. This effect is manifest as a cyclical change in climate with a period of about 41 kyr.

A further effect of planetary attraction is to cause the eccentricity of the Earth's orbit, at present 0.017, to change cyclically (Fig. 3.20). At one extreme of the cycle, the orbit is almost circular, with an eccentricity of only 0.005. The closest distance from the Sun at perihelion is then 99% of the furthest distance at aphelion. At the other extreme, the orbit is more elongate, although with an eccentricity of 0.058 it is only slightly elliptical. The perihelion distance is then 89% of the aphelion distance. These slight differences have climatic effects. When the orbit is almost circular, the insolation difference between summer and winter is negligible. However, when the orbit is most elongate, the insolation in winter is only 78% of the summer insolation. The cyclical variation in eccentricity has a dominant period of 405 kyr, which is caused by the attractions of Jupiter and Venus on the Earth's orbit. The huge mass of Jupiter makes this a very stable cycle and therefore it is very useful for dating ancient sedimentary records. The interplanetary attractions also generate periodicities of 95 kyr, 99 kyr, 124 kyr, and 131 kyr that combine to give a roughly 100 kyr cycle. Paleoclimatic records that cover a long enough interval of time often show fluctuations with periods around 100 kyr and 405 kyr.

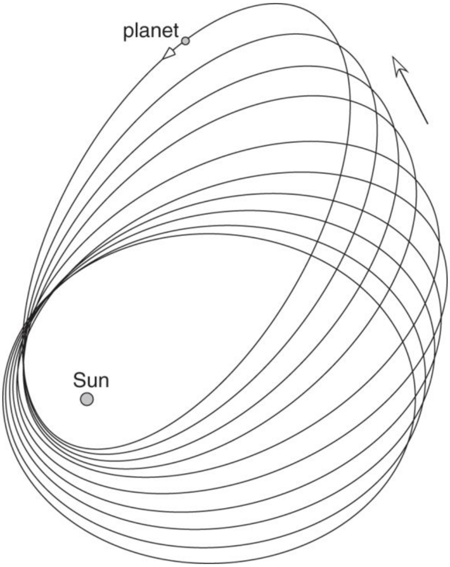


Fig. 3.20 Schematic illustration of the 100,000-year variations in eccentricity and rotation of the axis of the Earth's elliptical orbit. The effects are greatly exaggerated for ease of visualization.

Not only does planetary attraction cause the shape of the orbit to change, it also causes the perihelion–aphelion axis of the orbit to precess. The orbital ellipse is not truly closed, and the path of the Earth describes a rosette with a period that is also around 100 kyr (Fig. 3.20). The precession of perihelion interacts with the axial precession and modifies the observed period of the equinoxes. The 26 kyr axial precession is retrograde with a rate of 0.038 cycles/kyr; the 100 kyr orbital precession is prograde, which speeds up the effective precession rate to 0.048 cycles/kyr. This is equivalent to a retrograde precession with a period of about 21 kyr. A corresponding climatic fluctuation has been interpreted in many sedimentary deposits.

Climatic effects related to cyclical changes in the Earth's rotational and orbital parameters were first studied between 1920 and 1938 by a Yugoslavian astronomer, Milutin Milanković (anglicized to Milankovitch). Periodicities of 21 kyr, 41 kyr, 100 kyr, and 405 kyr – called the Milankovitch climatic cycles – modulate the Earth's insolation (Fig. 3.21) and have been observed in various sedimentary records ranging in age from Quaternary to Mesozoic. The recognition and measurement of cyclical features in sedimentary sections – a method called *cyclostratigraphy* – forms the basis of astrochronological dating of sediments in which Milankovitch cycles can be identified. Counting the cycles between securely dated stratigraphic levels has improved the dating of sedimentary sequences and the precision of the timescales for geological events and geomagnetic polarity reversals in the Mesozoic and Cenozoic eras, approximately the past 235 Myr.

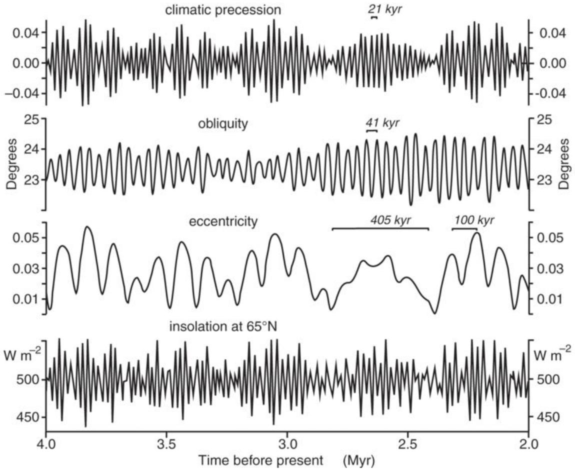


Fig. 3.21 Cyclical variations in precession index, obliquity, and eccentricity, and the insolation at 65°N; calculated for the time from 4 Myr to 2 Myr before present. Horizontal bars show the approximate lengths of the 21, 41, 100, and 405 kyr periodicities.

Source: Berger and Loutre, Astronomical theory of climate change, *J. Phys. IV France*, 121, 1–35, 2004.

Caution must be used in interpreting the cyclicities in older records, as the Milankovitch periods are dependent on astronomical parameters that change appreciably during the geological ages. However, the eccentricity period of 405 kyr results from the effects of Venus and Jupiter on the Earth's orbit, and the huge mass of Jupiter causes this period to be relatively stable. This makes it suitable for dating slowly deposited sedimentary rocks, such as pelagic limestones.

3.4 The Earth’s Figure and Gravity

3.4.1 The Figure of the Earth

The true surface of the Earth is uneven and irregular, partly land and partly water. For geophysical purposes the Earth's shape is represented by a smooth, closed surface, which is called the figure of the Earth. Early concepts of the figure were governed by religion, superstition and non-scientific beliefs. The first circumnavigation of the Earth, completed in 1522 by Ferdinand Magellan's crew, established that the Earth was definitely round. (The ancient Greek philosopher Eratosthenes had already inferred the roughly spherical shape of the Earth more indirectly.) Before the era of scientific awakening, the Earth's shape was believed to be a sphere. As confirmed by numerous photographs from spacecraft, this is in fact an excellent first approximation to Earth's shape that is adequate for solving many problems. The original suggestion that the Earth is a spheroid flattened at the poles is credited to Newton, who used a hydrostatic argument to account for the polar flattening. The slightly flattened shape permitted an explanation of why a clock that was precise in Paris lost time when placed near to the equator (see Section 3.1.2).

Earth's shape and gravity are intimately associated. The figure of the Earth is defined as the shape of an equipotential surface of gravity, in particular the one that coincides with mean sea level. The best mathematical approximation to the figure is an oblate ellipsoid, or spheroid (Fig. 3.22). The precise determination of the dimensions of the Earth (e.g., its polar and equatorial radii) is the main objective of the science of *geodesy*. It requires an exact knowledge of the Earth's gravity field, the description of which is the goal of *gravimetry*.

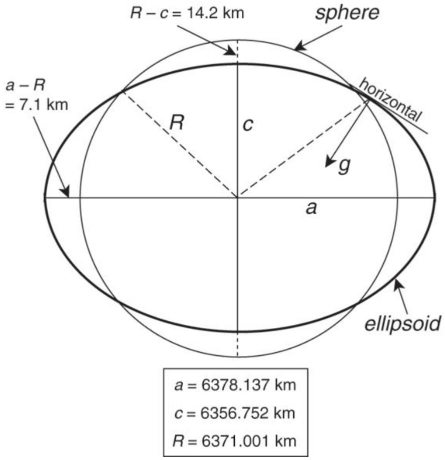


Fig. 3.22 Comparison of the dimensions of the *International Reference Ellipsoid* with a sphere of equal volume.

In addition to using land-based observations, modern analyses of the Earth's shape use precise observations of the orbits of artificial Earth satellites. These data are used to define a best-fitting oblate ellipsoid, called the *International Reference Ellipsoid*. In 1930 geodesists and geophysicists defined an optimum reference ellipsoid based on the best available data at the time. The dimensions of this figure have been subsequently refined as more exact data have become available. In 1980 the *International Association of Geodesy* adopted a *Geodetic Reference System* (GRS80) in which the reference ellipsoid has an equatorial radius (a) equal to 6378.137 km and a polar radius (c) equal to 6356.752 km. Subsequent determinations have resulted in only minor differences in the most important geodetic parameters. Some current values are listed in Table 3.1. The radius (R) of a sphere with the same volume as the ellipsoid is found from $R = (a^2 c)^{1/3}$ to be 6371.001 km. Compared to the best-fitting sphere, the spheroid is flattened by about 14.2 km at each pole and the equator bulges by about 7.1 km. The polar flattening f is defined as the ratio

Table 3.1 Fundamental parameters for Earth's shape, rotation and orbit

Parameter	Symbol	Value	Units

Physical parameter (CODATA 2014)			
Gravitational constant	G	$6.674\,08 \times 10^{-11}$	$\mathrm{m^3\,kg^{-1}\,s^{-2}}$
Terrestrial parameters (WGS 84), with CODATA 2014 value of G			
Geocentric gravitational constant	GE	$3.986\,004\,418 \times 10^{14}$	$\mathrm{m^3\,s^{-2}}$
Mass of the Earth: $E = (GE)/G$	E	$5.973\,328 \times 10^{24}$	kg
Earth’s equatorial radius	a	6,378,137.00	m
Earth’s polar radius: $c = a(1-f)$	c	6,356,752.3142	m
Radius of equivalent sphere: $R_0 = (a^2\,c)^{1/3}$	R_0	6,371,000.79	m
Theoretical gravity at equator	g_e	9.780 325 3359	$\mathrm{m\,s^{-2}}$
Theoretical gravity at poles	g_p	9.832 184 9378	$\mathrm{m\,s^{-2}}$
Mean angular velocity of rotation	Ω	$7.292\,115 \times 10^{-5}$	$\mathrm{rad\,s^{-1}}$
Dynamical form factor	J_2	$1.082\,629\,821 \times 10^{-3}$	
Flattening	f	1 : 298.257 223 563	
Equatorial acceleration ratio: $\mu = \Omega^2 a^2\,c/(GE)$	μ	1 : 289.873 010 407	

Dynamical ellipticity	H	1 : 305.441
Orbital parameters (IAU 2014)		
Astronomical unit	AU	149,597,870,700 m
Mass ratio: Sun to Earth	μ_{S}	332,946.0487
Mass ratio: Moon to Earth	μ_{L}	0.0123 000 371
Obliquity of the ecliptic (J2000)	ε_0	23° 26′ 21.406″
Obliquity of Moon's orbit to ecliptic		5° 8.7′
Eccentricity of Earth's orbit		0.0167 086
Eccentricity of Moon's orbit		0.0549

Sources: CODATA 2014 (Mohr et al., 2016); World Geodetic System 1984 (National Imaging and Mapping Agency, 2000); IAU 2014 (Luzum et al., 2011).

$$f = \frac{a - c}{a}. \tag{3.46}$$

The flattening of the optimum reference ellipsoid defined in 1930 was exactly 1/297. This ellipsoid, and the variation of gravity on its surface, served as the basis of gravimetric surveying for many years, until the era of satellite geodesy and highly sensitive gravimeters showed it to be too inexact. A recent best estimate of the flattening is $f = 3.352\,8107 \times 10^{-3}$ (i.e., $f = 1/298.257223563$).

If the Earth is assumed to be a rotating fluid in perfect hydrostatic equilibrium (as assumed by Newton's theory), the flattening should be 1/299.5, slightly smaller than the observed value. The hydrostatic condition assumes that the Earth has no internal strength. A possible explanation for the tiny discrepancy in f is that the Earth has sufficient strength to maintain a non-hydrostatic figure, and the present figure is inherited from a time of more rapid rotation. An additional contribution to the non-hydrostatic shape comes from internal density contrasts, related to convection in the Earth's mantle. This convective motion takes place over long time intervals and also results in a non-hydrostatic mass distribution.

The cause of the polar flattening is the deforming effect of the centrifugal acceleration. This is maximum at the equator, where the gravitational acceleration is smallest. Defining the parameter μ as the ratio of the equatorial centrifugal acceleration to the equatorial gravitational acceleration, we get:

$$\mu = \frac{\omega^2 a}{GE/a^2} = \frac{\omega^2 a^3}{GE}. \tag{3.47}$$

Inserting current values for the defining parameters (Table 3.1) in this formula gives a value of $\mu = 3.461392 \times 10^{-3}$ (i.e., $\mu = 1/288.901$). However, the exact formula for the variation of gravity with latitude (see Section 3.4.4.1) defines μ in a slightly different way, equivalent to the value of μ on the equator of a rotating *sphere* with equivalent volume to the Earth. This yields $\mu = 3.449\,787 \times 10^{-3}$ (i.e., $\mu = 1/289.873$).

As a result of the flattening, the distribution of mass within the Earth is not simply dependent on radius. The moments of inertia of the Earth about the rotation axis (C) and any axis in the equatorial plane (A) are unequal. As noted in the previous chapter, the inequality affects the way the Earth responds to external gravitational torques and is a determining factor in perturbations of the Earth's rotation. The principal moments of inertia define the *dynamical ellipticity*:

$$H = \frac{C - \frac{1}{2}(A+B)}{C} \approx \frac{C-A}{C}. \qquad (3.48)$$

The dynamical ellipticity is obtained from precise observations of the orbits of artificial satellites of the Earth (see Section 3.5). A modern determination of H is 3.2739515×10^{-3} (i.e., $H = 1/305.4413$).

3.4.2 Gravitational Potential of the Spheroidal Earth

The ellipsoidal shape changes the gravitational potential of the Earth from that of an undeformed sphere. In 1849 J. MacCullagh developed the following formula for the gravitational potential of any body at large distances from its center of mass:

$$U_{G} = G\frac{E}{r} - G\frac{(A+B+C-3I)}{2r^{3}}. \tag{3.49}$$

The first term decreases as r^{-1} and is the gravitational potential of a point mass or sphere with mass E (Eqs. 3.10 and 3.14); for the Earth it describes the potential of the undeformed globe. If the reference axes are centered on the body's center of mass, there is no term in r^{-2}. The second term decreases as r^{-3} and is due to deviations from the spherical shape. For the flattened Earth it results from the mass displacements due to the rotational deformation. The parameters A, B, and C are the principal moments of inertia of the body and I is the moment of inertia about the line OP joining the center of mass to the point of observation (Fig. 3.23). In order to express the potential accurately, an infinite number of terms of higher degree in r are needed. In the case of the Earth these can often be neglected, because the next term is about 1000 times smaller than the second term.

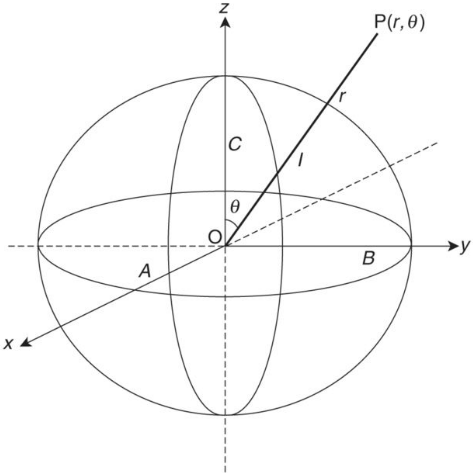


Fig. 3.23 Parameters of the ellipsoid used in MacCullagh's formula. *A*, *B*, and *C* are moments of inertia about the *x*-, *y*- and *z*-axes, respectively, and *I* is the moment of inertia about the line OP.

For a body with planes of symmetry, *I* is a simple combination of the principal moments of inertia. Setting *A* equal to *B* for rotational symmetry, and defining the angle between OP and the rotation axis to be *θ*, the expression for *I* is

$$I = A\sin^2\theta + C\cos^2\theta. \tag{3.50}$$

MacCullagh's formula for the ellipsoidal Earth then becomes

$$U_G = -G\frac{E}{r} + G\frac{(C-A)}{r^3}\frac{(3\cos^2\theta - 1)}{2}. \tag{3.51}$$

Box 3.3 Legendre Polynomials

In the triangle depicted in Fig. B3.3.1, the side *u* is related to the other two sides *r* and *R*, and the angle *θ* they enclose, by the cosine law. The expression for $1/u$ can then be written:

$$\frac{1}{u}=\frac{1}{(R^2+r^2-2rR\cos\theta)^{1/2}}=\frac{1}{R}\left[1+\left(\frac{r^2}{R^2}-2\frac{r}{R}\cos\theta\right)\right]^{-1/2},$$
(1)

which on expanding becomes

$$\frac{1}{u}=\frac{1}{R}\left[1+\frac{r}{R}\cos\theta+\frac{r^2}{R^2}\left(\frac{3\cos^2\theta-1}{2}\right)\right.$$
$$\left.+\frac{r^3}{R^3}\left(\frac{5\cos^3\theta-3\cos\theta}{2}\right)+\cdots\right].$$
(2)

This infinitely long series of terms in (r/R) is called the reciprocal distance formula. It can be written in shorthand form as

$$\frac{1}{u}=\frac{1}{R}\sum_{n=0}^{\infty}\left(\frac{r}{R}\right)^{n}P_n(\cos\theta).$$
(3)

The angle θ in this expression describes the angular deviation between the side r and the reference side R. The functions $Pn(\cos\theta)$ in the sum are called the *ordinary Legendre polynomials* of degree n in $\cos\theta$. They are named after a French mathematician Adrien Marie Legendre (1752–1833). Each polynomial is a coefficient of $(r/R)^n$ in the infinite sum of terms for ($1/u$), and so has degree n. Writing $\cos\theta = x$, and $P_n(\cos\theta) = P_\mathrm{n}(x)$, the first few polynomials, for $n = 0, 1, 2,$ and 3, respectively, are as follows

$$P_0(x)=1 \qquad\qquad P_2(x)=\frac{1}{2}(3x^2-1)$$
$$P_1(x)=x \qquad\qquad P_3(x)=\frac{1}{2}(5x^3-3x).$$
(4)

By substituting cosθ for x, these expressions can be converted into functions of cosθ. Legendre discovered that the polynomials satisfied the following second-degree differential equation, in which n is an integer and $y = P_n(x)$:

$$\frac{\partial}{\partial x}(1-x^2)\frac{\partial y}{\partial x}+n(n+1)y=0. \tag{5}$$

This, named in his honor, is the *Legendre equation*. It plays an important role in geophysical potential theory for situations expressed in spherical coordinates that have rotational symmetry about an axis. This is, for example, the case for the gravitational attraction of a spheroid, the simplified form of the Earth's shape.

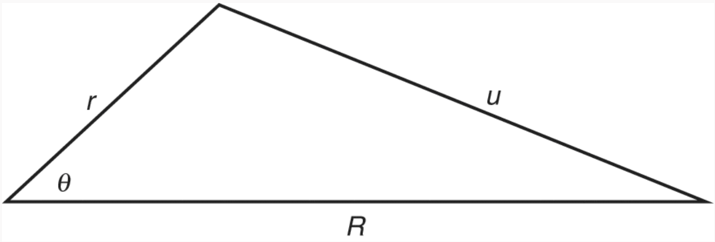


Fig. B3.3.1 Reference triangle for derivation of Legendre polynomials.

The derivation of an individual polynomial of degree n is rather tedious if the expanded expression for $(1/u)$ is used. A simple formula for calculating the Legendre polynomials for any degree n was developed by another French mathematician, Olinde Rodrigues (1794–1851). The *Rodrigues formula* is

$$P_n(x)=\frac{1}{2^n n!}\frac{\partial^n}{\partial x^n}(x^2-1)^n. \tag{6}$$

A relative of this equation encountered in many problems of potential theory is the *associated Legendre equation*, which written as a function of x is

$$\frac{\partial}{\partial x}(1-x^2)\frac{\partial y}{\partial x}+\left(n(n+1)-\frac{m^2}{(1-x^2)}\right)y=0.$$

(7)

The solutions of this equation involve two integers, the *degree n* and *order m*. As in the case of the ordinary Legendre equation, the solutions are polynomials in x, which are called the *associated Legendre polynomials* and written $P_n^m(x)$. A modification of the Rodrigues formula allows easy computation of these functions from the ordinary Legendre polynomials:

$$P_n^m(x)=(1-x^2)^{m/2}\frac{\partial^m}{\partial x^m}P_n(x).$$

(8)

To express the associated Legendre polynomials as functions of θ, i.e., as $P_n^m(\cos\theta)$, it is again only necessary to substitute $\cos\theta$ for x.

The function $(3\cos^2\theta - 1)/2$ is a second-degree polynomial in $\cos\theta$, written as $P_2(\cos\theta)$. It belongs to a family of functions called Legendre polynomials (Box 3.3). Using this notation, MacCullagh's formula for the gravitational potential of the oblate ellipsoid becomes

$$U_G=-G\frac{E}{r}+G\frac{(C-A)}{r^3}P_2(\cos\theta).$$

(3.52)

This can be written in the alternative form

$$U_G = -G\frac{E}{r}\left(1 - \left(\frac{C-A}{Er^2}\right)\left(\frac{R}{r}\right)^2 P_2(\cos\theta)\right). \tag{3.53}$$

Box 3.4 Spherical Harmonics

Many natural forces are directed toward a central point. Examples are the electrical field of a point charge, the magnetic field of a single magnetic pole, and the gravitational acceleration toward a center of mass. The French astronomer and mathematician Pierre-Simon, marquis de Laplace (1749–1827) showed that, in order to fulfill this basic physical condition, the potential of the field must satisfy a second-order differential equation, the *Laplace equation*. This is one of the most famous and important equations in physics and geophysics, as it applies to many situations in potential theory. For the gravitational potential U_G the Laplace equation is written in Cartesian coordinates (x, y, z) as

$$\frac{\partial^2 U_G}{\partial x^2} + \frac{\partial^2 U_G}{\partial y^2} + \frac{\partial^2 U_G}{\partial z^2} = 0. \tag{1}$$

In spherical polar coordinates (r, θ, ϕ), the Laplace equation becomes

$$\frac{1}{r^2}\frac{\partial}{\partial r}r^2\frac{\partial U_G}{\partial r} + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\sin\theta\frac{\partial U_G}{\partial\theta} + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 U_G}{\partial\phi^2} = 0. \tag{2}$$

The variation with azimuth ϕ disappears for symmetry about the rotational axis. The general solution of the Laplace equation for rotational symmetry (e.g., for a spheroidal Earth) is

$$U_G = \sum_{n=0}^{\infty} \left(A_n r^n + \frac{B_n}{r^{n+1}} \right) P_n(\cos\theta), \tag{3}$$

where $P_n(\cos\theta)$ is an ordinary Legendre polynomial of degree n and the coordinate θ is the angular deviation of the point of observation from the reference axis (see Box 3.3). In geographic coordinates θ is the colatitude.

If the potential field is not rotationally symmetric – as is the case, for example, for the geoid and the Earth's magnetic field – the solution of the Laplace equation varies with azimuth ϕ as well as with radius r and axial angle θ, and is given by

$$U_G = \sum_{n=0}^{\infty} \left(A_n r^n + \frac{B_n}{r^{n+1}} \right) \sum_{m=0}^{n} (a_n^m \cos m\phi + b_n^m \sin m\phi) P_n^m(\cos\theta), \tag{4}$$

where in this case $P_n^m(\cos\theta)$ is an associated Legendre polynomial of degree n and order m, as described in Box 3.3. This equation can in turn be written in modified form as

$$U_G = \sum_{n=0}^{\infty} \left(A_n r^n + \frac{B_n}{r^{n+1}} \right) \sum_{m=0}^{n} Y_n^m(\theta, \phi). \tag{5}$$

The function

$$Y_n^m(\theta, \phi) = (a_n^m \cos m\phi + b_n^m \sin m\phi)\, P_n^m(\cos\theta) \tag{6}$$

is called a *spherical harmonic function*, because it has the same value when θ or ϕ is increased by an integral multiple of 2π. It describes the variation of the potential with the coordinates θ

and ϕ on a spherical surface (i.e., for which r is a constant). Spherical harmonic functions are used, for example, for describing the variations of the gravitational and magnetic potentials, geoid height, and global heat flow with latitude and longitude on the surface of the Earth.

Terms in the potential expansion for which the degree n and order m of the coefficients are equal (e.g., a_1^1, b_2^2) are called *sectorial harmonics*. Their symmetry relative to the Earth’s axis is characterized by an even number of sectors around the equator in which the shape of the potential surface is alternately above and below the spherical shape (Fig. B3.4.1). The potential terms for which $m \neq n$ (e.g., a_2^1, b_3^2) are known as *tesseral harmonics*. Their pattern of symmetry is defined by the intersections of circles of latitude and longitude, which outline alternating domains in which the shape of the potential surface departs from that of a sphere.

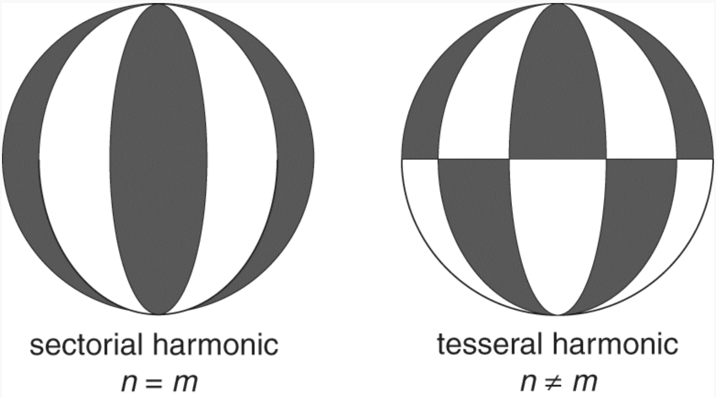


Fig. B3.4.1 Symmetry of sectorial and tesseral spherical harmonic functions.

By superposing many terms corresponding to these simple geometries, a field of great complexity can be generated.

Potential theory requires that the gravitational potential of the spheroidal Earth must satisfy an important equation, the Laplace equation (Box 3.4). The solution of this equation is the sum of an infinite number of terms of increasing degree in $1/r$, each involving an appropriate Legendre polynomial:

$$U_G = -G\frac{E}{r}\left(1 - \sum_{n=2}^{\infty}\left(\frac{R}{r}\right)^n J_n P_n(\cos\theta)\right). \tag{3.54}$$

In this equation the coefficients J_n multiplying $P_n(\cos\theta)$ determine the relative importance of the term of nth degree. The values of J_n are obtained from satellite geodesy: $J_2 = 1082.6 \times 10^{-6}$; $J_3 = -2.54 \times 10^{-6}$; $J_4 = -1.59 \times 10^{-6}$; higher degrees are insignificant. The most important coefficient is the second-degree one, the dynamical form-factor J_2, which describes the effect of the polar flattening on the Earth's gravitational potential. Comparison of terms in Eqs. (3.52) and (3.53) gives the result

$$J_2 = \frac{C - A}{ER^2}. \tag{3.55}$$

The term of next higher degree ($n = 3$) in Eq. (3.54) describes the deviations from the reference ellipsoid that correspond to a pear-shaped Earth (Fig. 3.24). These deviations are of the order of 7–17 m, 1000 times smaller than the deviations of the ellipsoid from a sphere, which are of the order of 7–14 km.

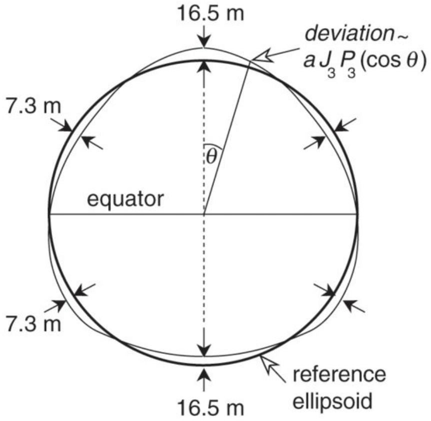

16.5 m
deviation ~ $aJ_3P_3(\cos\theta)$
7.3 m
θ
equator
7.3 m
16.5 m
reference ellipsoid

Fig. 3.24 The third-order term in the gravitational potential describes a pear-shaped Earth. The deviations from the reference ellipsoid are of the order of 7–17 m, much smaller than the deviations of the ellipsoid from a sphere, which are of the order of 7–14 km.

3.4.3 Gravity and Its Potential

The potential of gravity (U_g) is the sum of the gravitational and centrifugal potentials. It is often called the *geopotential*. At a point on the surface of the rotating spheroid it can be written

$$U_g = U_G - \frac{1}{2}\omega^2 r^2 \sin^2\theta. \tag{3.56}$$

If the free surface is an equipotential surface of gravity, then U_g is everywhere constant on it. The shape of the equipotential surface is constrained to be that of the spheroid with flattening f. Under these conditions a simple relation is found between f, μ, and J_2:

$$J_2 = \frac{1}{3}(2f - \mu). \tag{3.57}$$

By equating Eqs. (3.55) and (3.57) we obtain the following relationship:

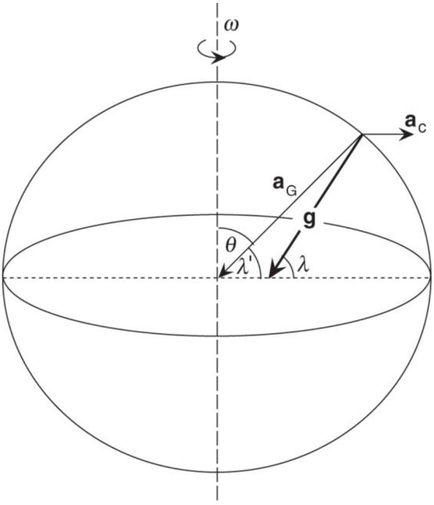
$$\frac{C - A}{ER^2} = \frac{1}{3}(2f - \mu). \tag{3.58}$$

This yields useful information about the variation of density within the Earth. The quantities f, μ, and $(C - A)/C$ are each equal to approximately 1/300. Inserting their values in the equation gives $C \approx 0.33\ ER^2$. Compare this value with the principal moments of inertia of a hollow spherical shell ($0.66\ ER^2$) and a solid sphere with uniform density ($0.4\ ER^2$). The concentration of mass near the center causes a reduction in the multiplying factor from 0.66 to 0.4. The value of 0.33 for the Earth implies that, in comparison with a uniform solid sphere, the density must increase toward the center of the Earth.

3.4.4 Normal Gravity

The direction of gravity at a point is defined as perpendicular to the equipotential surface through the point. This defines the *vertical* at the point, while the plane tangential to the equipotential surface defines the *horizontal* (Fig. 3.22). A consequence of the spheroidal shape of the Earth is that the vertical direction is generally not radial, except on the equator and at the poles.

On a spherical Earth there is no ambiguity in how we define latitude. It is the angle at the center of the Earth between the radius and the equator, the complement to the polar angle θ. This defines the geocentric latitude λ'. However, the geographic latitude in common use is not defined in this way. It may be found by geodetic measurement of the angle of elevation of a fixed star above the horizon. But the horizontal plane is tangential to the ellipsoid, not to a sphere (Fig. 3.22), and the vertical direction (i.e., the local direction of gravity) intersects the equator at an angle λ that is slightly larger than the geocentric latitude λ' (Fig. 3.25). The difference $(\lambda - \lambda')$ is zero at the equator and poles and reaches a maximum at a latitude of 45°, where it amounts to only 0.19° (about 12′).



gravity = $\mathbf{g} = \mathbf{a}_G + \mathbf{a}_c$

Fig. 3.25 Gravity on the ellipsoidal Earth is the vector sum of the gravitational and centrifugal accelerations and is not radial; consequently, geographic latitude (λ) is slightly larger than geocentric latitude (λ′).

The *International Reference Ellipsoid* is the standardized reference figure of the Earth. The theoretical value of gravity on the rotating ellipsoid can be computed by differentiating the gravity potential (Eq. 3.56). This yields the radial and transverse components of gravity, which are then combined to give the following formula for gravity normal to the ellipsoid:

$$g_n(\lambda) = g_e(1 + \beta_1 \sin^2\lambda + \beta_2 \sin^2 2\lambda), \tag{3.59}$$

where, to second order in f and μ,

$$\begin{aligned} g_e &= G\frac{E}{a^2}\left(1 + f - \frac{3}{2}\mu + f^2 - \frac{27}{14}f\mu\right) \\ \beta_1 &= \frac{5}{2}\mu - f + \frac{15}{4}\mu^2 - \frac{17}{14}f\mu \\ \beta_2 &= \frac{1}{8}f^2 - \frac{5}{8}f\mu. \end{aligned} \tag{3.60}$$

Equation (3.59) is known as the *normal gravity formula*. It was established in 1930 and the defining constants were progressively refined with time as knowledge of the Earth's shape and gravity field improved. The constants in the formula were defined in 1980 for the Geodetic Reference System (GRS80) to be: g_e = 9.780 327 m s^{-2}; β_1 = 5.30244 × 10^{-3}; β_2 = −5.8 × 10^{-6}. They allow calculation of normal gravity at any latitude with an accuracy of 0.1 mGal. Modern instruments can measure gravity differences with even greater precision, in which case a more exact formula, accurate to

0.0001 mGal, can be used. The normal gravity formula is very important in the analysis of gravity measurements on the Earth, because it gives the theoretical variation of normal gravity (g_n) with latitude on the surface of the reference ellipsoid.

The normal gravity formula gives $g_p = 9.832\,186$ m s^{-2}. Numerically, this gives an increase in gravity from equator to pole of approximately 5.186×10^{-2} m s^{-2}, or 5186 mGal. There are two causes of the poleward increase of gravity. First, the distance to the center of mass of the Earth is shorter at the poles than at the equator. This gives a stronger gravitational acceleration (a_G) at the poles. The difference is

$$\Delta a_G = \left(\frac{GE}{c^2} - \frac{GE}{a^2}\right). \tag{3.61}$$

This gives an excess gravity of approximately 6600 mGal at the poles.

The second cause is the variation with latitude of the centrifugal force. Its effect in diminishing gravity is largest at the equator, where it equals μa_G, and is zero at the poles. This results in a poleward increase of gravity of about 3375 mGal. Taken together, these figures indicate that gravity should increase by a total of 9975 mGal from equator to pole, instead of the observed difference of 5186 mGal. The discrepancy can be resolved by taking into account a third factor. The computation of the difference in gravitational attraction is not so simple as indicated by Eq. (3.61), but must take into account the effect of the equatorial bulge. It places an excess of mass under the equator, increasing the equatorial gravitational attraction and thereby reducing the gravity decrease from equator to pole.

3.4.4.1 International Gravity Formula

The parameters defining the reference ellipsoid have been refined progressively, as a consequence of improvements in instrumentation, computer, and satellite technology. The development of satellite geodesy (Section 3.5) greatly improved the quantity and quality of global gravimetric and geodetic data. For example, the parameters of the Earth Gravitational Model 1996 were derived for a spherical harmonic analysis to degree and order 360; by contrast, the *Earth Gravitational Model 2008* (Pavlis et al., 2012) is computed to degree and order 2159, involving approximately 4.7 million coefficients. At the same time, improvements in instrumentation have made it possible to measure gravity with very high precision; the superconducting gravimeter has a sensitivity of one nanogal (10^{-11} m s^{-2}). Equation (3.59) became inadequate for calculating theoretical gravity at a point and it was replaced as a global reference by a new version.

The parameters g_e, β_1, and β_2 in Eq. (3.59) are defined only to second order in f and μ. Truncation of the defining equations at higher order is possible, but the formula becomes unwieldy and is not sufficiently accurate for modern gravimetry. Consequently, this approximate formula for theoretical gravity has been replaced by an exact formula that involves no truncation. This equation was derived by Somigliana (1929) and adopted for the *World Geodetic System 1984* (WGS84). It defines normal gravity g_n at geodetic latitude λ as

$$g_n(\lambda) = \frac{a g_e \cos^2\lambda + c g_p \sin^2\lambda}{\sqrt{a^2 \cos^2\lambda + c^2 \sin^2\lambda}}.$$

(3.62)

Here, a and c are the semi-major and semi-minor axes of the ellipsoid, and g_e and g_p are the values of theoretical gravity at the equator and pole, respectively. The defining parameters are known very accurately from satellite geodesy. The WGS84 values are: g_e = 9.7803253359 m s^{-2}, g_p = 9.8321849378 m s^{-2}, a = 6,378,137 m, c = 6,356,752.314 m; flattening f = 1/298.257223563. Equation (3.61) can be written in the form

$$g_n(\lambda) = g_e \frac{(1 + k \sin^2\lambda)}{\sqrt{1 - e^2 \sin^2\lambda}}, \tag{3.63}$$

where

$$k = \frac{cg_p - ag_e}{ag_e} \qquad \text{and} \qquad e^2 = \frac{a^2 - c^2}{a^2}. \tag{3.64}$$

The quantity e^2 is the squared eccentricity of the ellipsoid. The values of these parameters are $k = 0.00193185265241$ and $e^2 = 0.00669437999014$. This formula replaces the earlier normal gravity formula (Eq. 3.59) and allows calculation of the theoretical gravity at a given latitude λ with a precision of better than a microgal.

The adoption of Eq. (3.63) for normal gravity affects the definition of the acceleration ratio μ, which in WGS84 becomes

$$\mu = \frac{\omega^2 a^2 c}{GE}. \tag{3.65}$$

This is analogous to the value of μ for a rotating spherical Earth with radius R and equivalent volume to the ellipsoid, for which $R^3 = (a^2 c)$. The corresponding value of μ is 0.00344978650684, i.e., about 1/289.873.

3.4.4.2 Clairaut's Theorem

In Eq. (3.60) the second-order terms f^2, μ^2, and $f\mu$ are about 300 times smaller than the first-order terms f and μ, and thus the constant β_2 in Eq. (3.59) is about 1000 times smaller than β_1. If we drop second-order terms and use $\lambda = 90°$, the value of normal gravity at the pole is $g_p = g_e(1 + \beta_1)$, so by rearranging and retaining only first-order terms, we get

$$\frac{g_p - g_e}{g_e} = \frac{5}{2}\mu - f. \tag{3.66}$$

This expression is called Clairaut's theorem. It was developed in 1743 by a French mathematician, Alexis-Claude Clairaut, who was the first to relate the variation of gravity on the rotating Earth with the flattening of the spheroid.

3.4.5 The Geoid

The International Reference Ellipsoid is a close approximation to the equipotential surface of gravity, but it is really a mathematical convenience. The physical equipotential surface of gravity that best fits global mean sea level is called the *geoid*. It reflects the true distribution of mass inside the Earth and differs from the theoretical ellipsoid by small amounts. Far from land the geoid agrees with the free ocean surface, excluding the temporary perturbing effects of tides, currents and winds. Over the continents the geoid is affected by the mass of land above mean sea level (Fig. 3.26a). The mass within the ellipsoid causes a downward gravitational attraction toward the center of the Earth, but a hill or mountain whose center of gravity is outside the ellipsoid causes an upward attraction. This causes a local elevation of the geoid above the ellipsoid. The displacement between the geoid and the ellipsoid is called a *geoid undulation*; the elevation caused by the mass above the ellipsoid is a positive undulation. Additional geoid undulations, called *dynamic topography*, are caused by very long term up- and downward motions related to convection in the Earth's mantle.

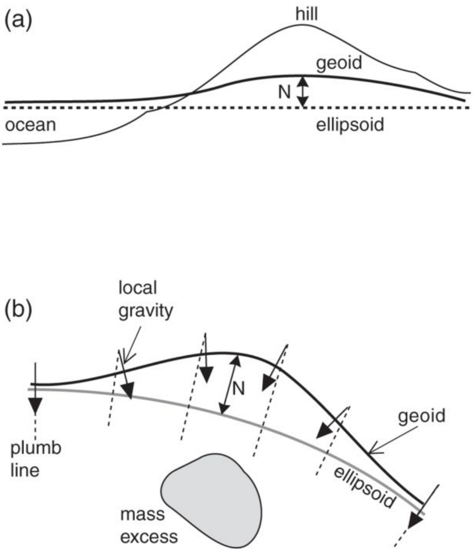


Fig. 3.26 (a) A mass outside the ellipsoid or (b) a mass excess below the ellipsoid elevates the geoid above the ellipsoid. N is the geoid undulation.

3.4.5.1 Geoid Undulations

In computing the theoretical figure of the Earth the distribution of mass beneath the ellipsoid is assumed to be homogeneous. A local excess of mass under the ellipsoid will deflect and strengthen gravity locally. The potential of the ellipsoid is achieved further from the center of the Earth. The equipotential surface is forced to warp upward while remaining normal to gravity. This gives a positive geoid undulation over a mass excess under the ellipsoid (Fig. 3.26b). Conversely, a mass deficit beneath the ellipsoid will deflect the geoid below the ellipsoid, causing a negative geoid undulation. As a result of the uneven topography and heterogeneous internal mass distribution of the Earth, the geoid is a bumpy equipotential surface.

The potential of the geoid is represented mathematically by *spherical harmonic functions* that involve the *associated Legendre polynomials* (Box 3.4). These are more complicated than the ordinary Legendre polynomials used to describe the gravitational potential of the ellipsoid (Eq. 3.54). Until now we have only considered variation of the potential with distance r and with the colatitude angle θ. This is an oversimplification, because density variations within the Earth are not symmetrical about the rotation axis. The geoid is an equipotential surface for the real density distribution in the Earth, and so the potential of the geoid varies with longitude as well as colatitude. These variations are taken into account by expressing the potential as a sum of spherical harmonic functions, as described in Box 3.4. This representation of the geopotential is analogous to the simpler expression for the gravitational potential of the rotationally symmetric Earth using a series of Legendre polynomials.

In modern analyses the coefficient of each term in the geopotential – similar to the coefficients J_n in Eq. (3.54) – can be calculated up to a high harmonic degree. The terms up to a selected degree are then used to compute a model of the geoid and the Earth's gravity field. A combination of satellite data and surface gravity measurements was used to construct the Goddard Earth Model (GEM) 10. A global comparison between a reference ellipsoid with flattening 1/298.257 and the geoid surface computed from the GEM 10 model shows long-wavelength geoid undulations (Fig. 3.27). The largest negative undulation (–105 m) is in the Indian Ocean south of India, and the largest positive undulation (+73 m) is in the equatorial Pacific Ocean north of Australia. These large-scale features are too broad to be ascribed to shallow crustal or lithospheric mass anomalies. They can be explained by density variations in the lower mantle and dynamic topography.

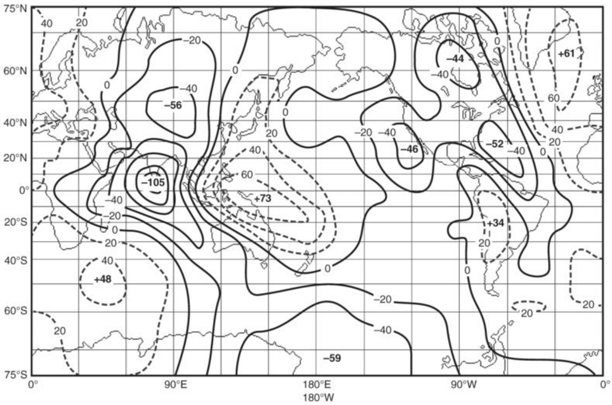


Fig. 3.27 World map of geoid undulations relative to a reference ellipsoid of flattening $f = 1/298.257$.

Source: F.J. Lerch, S.M. Klosko, R.E. Laubscher and C.A. Wagner, Gravity model improvement using Geos 3 (GEM 8 and 10), *J. Geophys. Res.* 84, 3897–3916, 1979.

3.5 Space Geodesy

3.5.1 Geodesy with Artificial Satellites

Since the early 1960s knowledge of the geoid has been dramatically enhanced by the development of space geodesy. This consists of techniques for observing the geoid and gravitational field by using sources and methods that are employed or originate outside the Earth. Satellite geodesy encompasses many of these methods, which involve observing artificial satellites from the Earth and observing the Earth with satellite-borne instruments. Additionally, the space-based method of VLBI uses signals from extra-galactic sources to study the Earth's rotation and interpret plate tectonic motions.

The motions of artificial satellites that orbit the Earth are influenced by the Earth's mass distribution. The most important interaction is the simple balance between the centrifugal force and the gravitational attraction of the Earth's mass, which determines the radius of the satellite's orbit. Analysis of the precession of the Earth's rotation axis (Section 3.3.5.4) shows that it is determined by the dynamical ellipticity H, which depends on the difference between the principal moments of inertia resulting from the rotational flattening. In principle, the gravitational attraction of an artificial satellite on the Earth's equatorial bulge also contributes to the precession, but the effect is too tiny to be measurable. However, the inverse attraction of the equatorial bulge on the satellite causes the orbit of the satellite to precess around the rotation axis. The plane of the orbit intersects the equatorial plane in the *line of nodes*. Let this be represented by the line CN_1 in Fig. 3.28. On the next passage of the satellite around the Earth the precession of the orbit has moved the nodal line to a new position CN_2. The orbital precession in this case is *retrograde*; the nodal line *regresses*. For a satellite orbiting in the same sense as the Earth's rotation the longitude of the nodal line shifts gradually westward; if the orbital sense is opposite to the Earth's rotation, the longitude of the nodal line shifts gradually eastward. Because of the precession of its orbit, the path of a satellite eventually covers the entire Earth between the north and south circles of latitude defined by the inclination of the orbit. The profusion of high-quality satellite data is the best source for calculating the dynamical ellipticity or the related parameter J_2 in the gravity potential. Observations of satellite orbits are so precise that small perturbations of the orbit can be related to the gravitational field and to the geoid. In fact, the positions of the satellites of the Global Positioning System (GPS) are known to within centimeters.

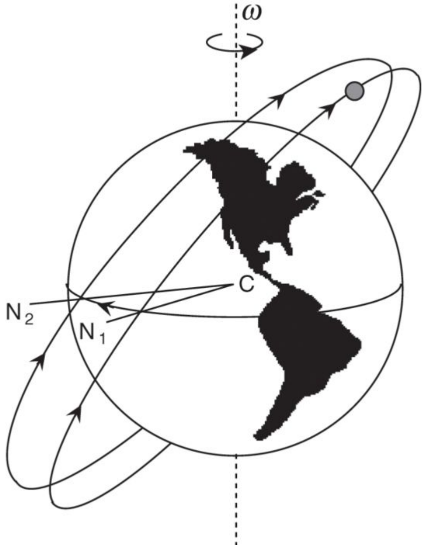


Fig. 3.28 The retrograde precession of a satellite orbit causes the line of nodes (CN_1, CN_2) to change position on successive equatorial crossings.

3.5.2 Satellite Laser-Ranging

The accurate tracking of a satellite orbit can be achieved by satellite laser-ranging (SLR). The spherical surface of the target satellite is covered with numerous retroreflectors. A retroreflector consists of three orthogonal mirrors that form the corner of a cube; it reflects an incident beam of light back along its path (Box 1.4). A brief pulse of laser light with a wavelength of 532 nm is sent from the tracking station on Earth to the satellite, and the two-way travel time of the reflected pulse is measured. Using the accurately known speed of light, the distance of the satellite from the tracking station is obtained. The accuracy of a single range measurement is about 1 cm.

The Laser Geodynamics Satellites (LAGEOS) and France’s Starlette and Stella satellites have been tracked for many years. LAGEOS-1, launched in 1976, has a very high orbit at 5858–5958 km altitude. The inclination of the orbit is 109.8° (i.e., its orbital sense is opposite to the Earth’s rotation), and the nodal line of the orbit advances at 0.343° per day. The orbit of LAGEOS-2, launched in 1992, has an inclination of 52.65°. Each LAGEOS satellite is passive (there are no on-board electronics) and its spherical surface is covered with 426 cube-corner retroreflectors (Box 1.4). Starlette, launched in 1975, orbits at an altitude of 806–1108 km; its orbit is inclined at 50° and its nodal line regresses at 3.95° per day. Stella, launched in 1993, has an orbit at 800 km altitude, inclined at 98.6°; its polar orbit covers regions missed by Starlette. Each of the French satellites carries 60 cube-corner reflectors. The LAGEOS satellites have very stable orbits so they can be used to determine positions on Earth with great accuracy. They are tracked continuously by the global satellite laser-ranging network.

The track of a satellite is perturbed by many factors, including the Earth’s gravity field, solar and lunar tidal effects, and atmospheric drag. The perturbing influences of these factors can be computed and allowed for. The high accuracy of SLR results makes it possible to detect variations in the coordinates of the tracking stations. The motion of the Earth's pole of rotation can be deduced and the history of changes in position of the tracking station can be obtained. LAGEOS satellites have been tracked by numerous laser-tracking stations on five tectonic plates. The relative changes in position between pairs of stations can be compared with the rates of plate tectonic motion deduced from marine geophysical data. For example, a profile from the Yaragadee tracking station in Australia and the tracking station in Hawaii crosses the converging plate boundary between the Indo-Australian and Pacific plates (Fig. 3.29). The results of four years of measurement show a decrease of the arc distance between the two stations at a rate of 63 ± 3 mm yr^{-1}. This is in good agreement with the corresponding rate of 67 mm yr^{-1} inferred from paleomagnetic measurements for the relative rotation of the tectonic plates.

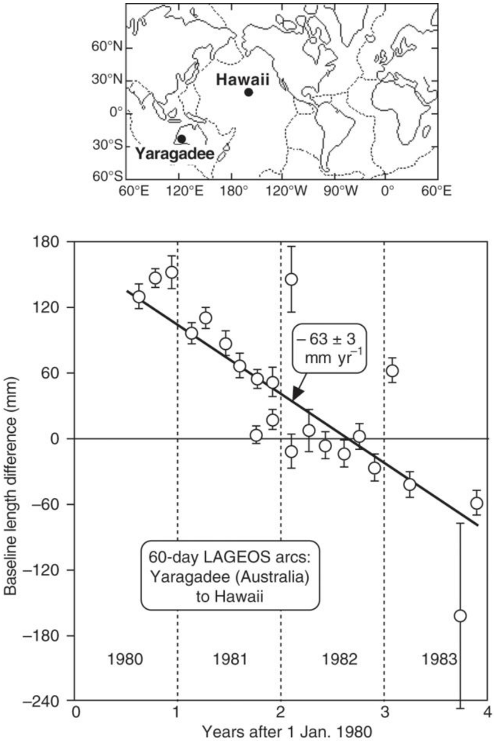


Fig. 3.29 Changes in the arc distance between satellite laser-ranging (SLR) stations in Australia and Hawaii determined from LAGEOS observations over a period of four years. The mean rate of

convergence, 63 ± 3 mm yr^{-1}, agrees well with the rate of 67 mm yr^{-1} deduced from plate tectonics.

Source: B.D. Tapley, B.E. Schutz and R.J. Eanes, Station coordinates, baselines and earth rotation from LAGEOS laser ranging: 1976–1984, *J. Geophys. Res.* 90, 9235–9248, 1985.

3.5.3 Satellite Altimetry

The position of an orbiting satellite can be located very accurately through a combination of ground-based SLR and on-board navigation instruments. The altitude of the satellite relative to the reference ellipsoid can be measured with a precision in the centimeter range. In satellite altimetry (Fig. 3.30), the tracked satellite carries a transmitter and receiver of microwave (radar) signals. A brief radar pulse is emitted from the spacecraft and reflected from the surface of the Earth. The two-way travel time is converted using the speed of light to a measurement of the height of the satellite above the Earth's surface. The difference between this height and the satellite's height above the ellipsoid gives the height of the topography relative to the reference ellipsoid. The precision over land areas is poorer than over the oceans, but over smooth land features like deserts and inland water bodies an accuracy of around a centimeter is achievable.

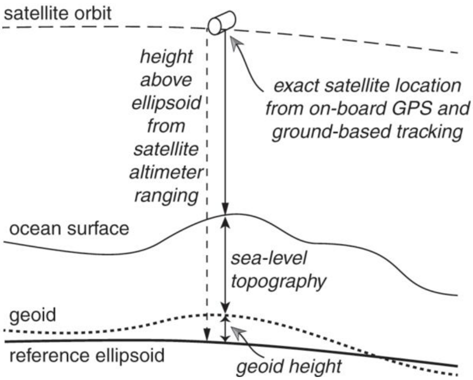

satellite orbit
height above ellipsoid from satellite altimeter ranging
exact satellite location from on-board GPS and ground-based tracking
ocean surface
sea-level topography
geoid
reference ellipsoid
geoid height

Fig. 3.30 In satellite altimetry over oceans the height of a satellite above the sea surface is measured by an on-board radar altimeter. The height above the reference ellipsoid is found by exact ground-based tracking of the satellite and by on-board GPS. The difference in heights is the level of the ocean surface above the ellipsoid. The difference between this sea level and the geoid defined by gravity measurements is the ocean surface topography.

Satellite altimeters are best suited for marine surveys, where high accuracy is possible. The satellite GEOS-3 flew from 1975 to 1978, SEASAT was launched in 1978, and GEOSAT was launched in 1985. Specifically designed for marine geophysical studies, these satellite altimeters revealed remarkable aspects of the marine geoid. The long-wavelength geoid undulations (Fig. 3.27) extend horizontally for hundreds of kilometers and have vertical amplitudes of up to several tens of meters. They are maintained by mantle-wide convection. The short-wavelength features can be accentuated by removing the computed geoid elevation up to a known degree and order. The data are presented on maps by suitable shading or coloring that emphasizes the elevated and depressed areas of the sea surface (Fig. 3.31).

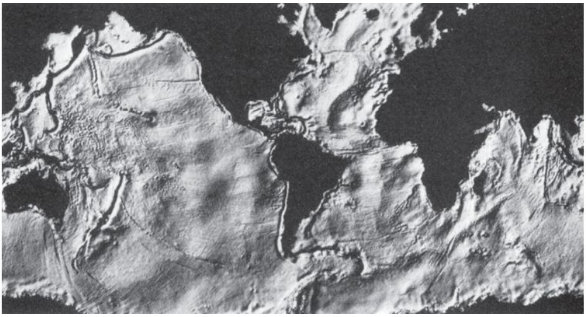


Fig. 3.31 The mean sea surface as determined from SEASAT and GEOS-3 satellite altimetry, after removal of long-wavelength features of the GEM-10B geoid up to order and degree 12. The surface is portrayed as though illuminated from the northwest.

Source: J.G. Marsh, C.J. Koblinsky, H.J. Zwally, A.C. Brenner and B.D. Beckley, A global mean sea surface based upon GEOS 3 and Seasat altimeter data, *J. Geophys. Res.* 97, 4915–4921, 1992.

There is a strong correlation between the short-wavelength anomalies in elevation of the mean sea surface and features of the sea-floor topography. Over the ocean ridge systems and seamount chains the mean sea surface (geoid) is raised. The locations of fracture zones, in which one side is elevated relative to the other, are clearly discernible. Very dark areas mark the locations of deep-ocean trenches, because the mass deficiency in a trench depresses the geoid.

Seaward of the deep-ocean trenches the mean sea surface is raised as a result of the upward flexure of the lithosphere before it plunges downward in a subduction zone.

3.5.4 Global Navigation Satellite Systems

Geodesy, the science of determining the three-dimensional coordinates of a position on the surface of the Earth, received an important boost with the advent of the satellite era. The first global navigation satellite system (GNSS), the US Navy Navigation Satellite System known as TRANSIT, consisted of six satellites in polar orbits about 1100 km above the surface of the Earth. Signals transmitted from these satellites were combined in a receiver on Earth with a signal generated at the same frequency in the receiver. Because of the motion of the satellite, the frequency of its signal was modified by the Doppler effect and was thus slightly different from the receiver-generated signal, producing a beat frequency. Using the speed of light, the beat signal was converted to the oblique distance between the satellite and receiver. By integrating the beat signal over a chosen time interval, the change in range to the satellite in the interval was obtained. This was repeated several times. The orbit of the satellite was known precisely from tracking stations on the ground, and so the position of the receiver could be calculated. Originally developed to support ballistic missile submarines in the 1960s, the system was extended to civilian navigation purposes, especially for fixing the position of a ship at sea. The TRANSIT program was terminated in 1996, and succeeded by the more precise Global Positioning System.

The Navigation Satellite Timing and Ranging Global Positioning System (NAVSTAR GPS or, more commonly, just GPS) utilizes satellites in much higher orbits, at an altitude of around 20,200 km (i.e., a radial distance of 26,570 km), with an orbital period of half a sidereal day. The GPS system consists of more than 30 satellites. They are distributed in six orbital planes, equally spaced at 60° intervals around the equator and inclined to the equator at about 55°. Between five and eight GPS satellites are visible at any time and location on Earth. Each satellite broadcasts its own predetermined position and reference signal every six seconds. The time difference between emission and reception on Earth gives the "pseudo-range" of the satellite, so-called because it must be corrected for errors in the clock of the receiver and for tropospheric refraction. Pseudo-range measurements to four or more satellites with known positions allows computation of the clock error and the position of the receiver within a few meters or better.

The American GPS satellites are part of a larger GNSS, in which similar constellations of satellites are deployed in Russia's GLONASS, China's BeiDou, and the European Union's Galileo system. In common usage, the term GPS from the original American system is often used indiscriminately to refer to positions located by any of these satellite systems.

The precision with which a position can be located by GPS depends on the quality of the receiver and signal processing. Low-cost single civilian-quality receivers (like those in mobile phones) have a positioning accuracy on the order of 5–10 m. Dual-frequency receivers can improve location to within a few centimeters. In scientific and military missions a roving receiver is used in conjunction with a base station (a fixed receiver), and differential signal processing improves the accuracy of location to around 1 cm. Long-term GPS observations at a fixed location augment the accuracy to the millimeter level by using carrier-phase tracking, which counts the total number of wavelengths plus any fractional wavelength in the received GPS signal. This method is accurate enough to measure plate tectonic motions of a few millimeters per year, as in the East African Rift Zone.

The use of GPS observations to study regional tectonics is illustrated by an investigation in the eastern Mediterranean. Repeated GPS surveys, augmented by continuous GPS measurements at networks of fixed stations, were carried out in 1988–1997 at 189 sites covering a region from western Greece to western Arabia. Horizontal crustal velocities were obtained for locations in Anatolia and on the African, Arabian, and Eurasian plates. The site mean velocities relative to the Eurasian plate show coherent anticlockwise rotation of the

Anatolian plate (Fig. 3.32). The largest velocities are in the central and southern Aegean, which are moving southwestward at 30 mm yr–^{1} relative to Eurasia. Anatolia is decoupled from Eurasia by the right-lateral, strike–slip North Anatolian fault, on which a mean slip rate of up to 24 mm yr–^{1} was determined.

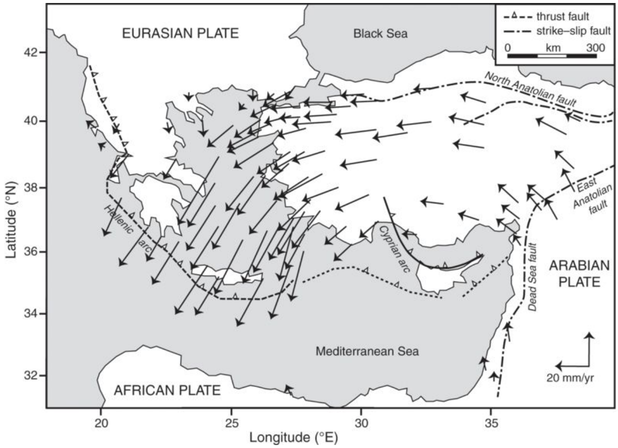

EURASIAN PLATE
Black Sea
thrust fault
strike–slip fault
0
km
300
North Anatolian fault
East Anatolian fault
Hellenic arc
Cyprian arc
Dead Sea fault
ARABIAN PLATE
Mediterranean Sea
AFRICAN PLATE
20 mm/yr
Latitude (°N)
42
40
38
36
34
32
20
25
30
35
Longitude (°E)

Fig. 3.32 Annual velocities of crustal motions in the eastern Mediterranean and Anatolia from GPS surveys in 1988–1997, plotted on a simplified tectonic map of the region. Dashed lines with triangles are thrust faults, with triangles on the overriding block. The heads of the velocity arrows are larger than the measurement errors.

Source: McClusky et al., 2000, Global Positioning System constraints on kinematics and dynamics in the eastern Mediterranean and Caucusus, *J. Geophys. Res.* 105, 569–5719, 2000.

3.5.5 Measurement of Gravity and the Geoid from Orbiting Satellites

The equipotential surface of gravity, the geoid (Section 3.4.5), is characterized by undulations caused by inhomogeneous distribution of mass in the Earth. Until recently, construction of a global model of the geoid was very laborious as it required combining data from many different sources of variable precision. Surface gravity measurements made on land or at sea were augmented by data from a large number of Earth-orbiting satellites. The resulting figure showed large-scale features (Fig. 3.27), but fine details were impossible to define accurately. Satellites in comparatively low orbits, a few hundreds of kilometers above the Earth's surface, can now be used in conjunction with land-based data and the GPS satellites orbiting at high altitudes (20,200 km) to measure the global gravity field and geoid with a precision that is several orders of magnitude better than was previously possible.

In 2000 the German CHAMP (***CHA***llenging ***M***ini-satellite ***P***ayload) satellite was inserted into a nearly circular, almost polar orbit with an initial altitude of 450 km. At this altitude the thin atmosphere is still capable of exerting drag, which lowers the altitude of the satellite to about 300 km over a five-year interval. Sensitive accelerometers (Box 4.2) on board the satellite allow correction for non-gravitational forces, such as atmospheric drag or the pressure of solar radiation. A highly precise GPS receiver on board the CHAMP satellite, using position data from up to 12 GPS satellites simultaneously, allows retrieval of CHAMP's position with an accuracy of a few centimeters. Whereas the orbits of earlier satellites were compiled from many comparatively short tracks measured when the satellite was in view of different ground stations, the CHAMP orbit is continuously tracked by the GPS satellites. Small perturbations of CHAMP's orbit may be tracked and modeled. The models of the Earth's gravity field and of the global geoid derived from CHAMP data were greatly improved in accuracy and definition over previous models.

The ***G***ravity ***R***ecovery ***A***nd ***C***limate ***E***xperiment (GRACE), a joint American–German project, was launched in 2002 and operated continuously for 15 years, three times longer than planned. The GRACE mission used two nearly identical satellites in near-circular polar orbits (inclination 89.5° to the equator), initially about 500 km above the Earth's surface. The twin satellites each carried on-board GPS receivers, which allowed precise determination of their absolute positions over the Earth at any time. The satellites traveled in tandem in the same orbital plane, separated by approximately 220 km along their track. Changes in gravity along the orbit were determined by observing small differences in the separation of the two satellites. This was achieved by using a highly accurate microwave ranging system. Each satellite carried a microwave antenna transmitting in the K-band frequency range (wavelength $\sim$ 1 cm) and directed accurately at the other satellite. With this ranging system the separation of the two satellites could be measured with a precision of one micrometer (1 μm). The GRACE measurement technology was employed in 2012 by the GRAIL mission to measure the gravitational field of the Moon.

The GRACE-type missions use an innovative method to measure gravity. As the satellite-pair orbits the Earth, it traverses variations in the gravity field due to the inhomogeneous mass distribution in the Earth. A mass excess causes the equipotential surface to bulge upward, and gravity is enhanced locally. The leading satellite encounters this anomaly first and is accelerated away from the trailing satellite, changing their separation. After the mass anomaly has been crossed, its mass causes the opposite effect, slowing down the leading satellite. Similar changes take place when the second satellite crosses the mass anomaly. The tiny changes in separation between the two satellites as they move along-track are detected by the accurate microwave ranging system. In conjunction with the exact location of each satellite by the on-board GPS devices, the GRACE measurement technology provides fine-scale definition of the gravity field and determination of the geoid. The GRACE mission ended in 2017. To provide continuity, a new mission, called GRACE-FO (where FO stands for "follow on"), was launched in 2018. It uses a laser interferometer to measure the satellite separation with even greater accuracy and map the Earth's gravitational field in finer detail than has yet been possible.

Prior to the GRACE mission, numerous field campaigns had amassed millions of gravity measurements worldwide. However, the density of global coverage was very uneven. Regions of the Earth that are difficult to access were sparsely covered. These included the polar regions, high mountain chains, jungles, deserts, and large oceanic areas. The quality of the gravity measurements was also uneven. The GRACE mission provided high-quality gravity data with uniform global coverage, which were combined with the terrestrial gravity data to form the Earth Gravity Model 2008 (EGM2008). The huge dataset was expressed in a set of spherical harmonic coefficients (Box 3.4) up to degree and order 2190. As a result, EGM2008 describes the gravity field in higher detail than any previous Earth model.

The first satellite investigation employing a high-sensitivity gradiometer (Section 4.1.3.2) to map changes in the Earth's gravitational field was the ***G***ravity field and steady-state ***O***cean ***C***irculation ***E***xplorer (GOCE) mission, launched by the European Space Agency in 2009. The satellite orbit had a high inclination of 96.7°, which provided almost complete global coverage, except for small areas around the poles. Gradiometer technology was employed because this gives greater resolution of more localized features of the Earth's mass distribution. The GOCE gradiometer consisted of three orthogonal arms, oriented respectively along the flight direction of the satellite, to the right of this track, and vertically down. At the opposite ends of each arm, 50 cm apart, a three-axis accelerometer (Box 4.2) was mounted. With this configuration the gradient of gravity in all directions could be measured, resolving geoid and gravity anomalies with high accuracy.

The GOCE mission resulted in a gravity model expressed by spherical harmonic coefficients up to degree and order 250, which involves calculating about 63,000 coefficients. The GOCE geoid model has the highest resolution to date, measuring geoid heights with an elevation accuracy of around 2 cm and gravity anomalies with an accuracy of 1 mGal. Where good terrestrial data are available, the results agree well with existing data and over uncharted regions they represent a significant improvement. For example, GOCE opened Antarctica to more detailed geophysical investigations than was previously possible. Except for an area with radius 700 km around the pole, the GOCE results describe the gravity field over most of the Antarctic continent, where gravity data were previously available only from coastal regions and from the GRACE mission. The GOCE data agree well with known topographic and tectonic features, opening the way to further tectonic analysis of the Antarctic continent, hitherto poorly known because of its thick ice cover.

A common feature of all investigations of the geoid is that its large-scale features do not correlate with the positions of continental or oceanic crustal blocks. The excess mass of the continents and the missing mass in the oceanic basins are compensated by the mechanism of isostasy (Section 4.3) and the origin of the long-wavelength geoid undulations must be deeper. They are associated with the effects of mantle convection (Section 9.9).

3.5.6 Time-Dependent Changes in the Earth’s Mass Distribution

The numerous high-quality measurements by satellite-borne instruments define the geoid and gravitational field in fine detail. The mathematical description in spherical harmonics (Box 3.4) has been determined to a high degree and order, permitting the definition of geoid features with a resolution of only a few hundred kilometers. An important feature of satellite observations is that they measure the gravitational field completely in about 30 days. This makes it possible to compare repeated surveys of a region, revealing time-dependent changes that accompany steady or transient exchanges of mass in the observed region. In fact, the satellite missions GRACE and GRACE-FO were specifically designed for this purpose. For example, the glacial isostatic adjustment (GIA, often termed post-glaciation isostatic rebound in older literature) of large regions such as northern Canada involves a redistribution of mass (topographic uplift, erosion). A similar effect accompanies the transfer of mass from the continents to the oceans in the global rise in sea level. In each case secular changes in the geoid result.

Satellite geodesy is able to monitor sea level very well. The French–American TOPEX/Poseidon satellite was launched to measure ocean surface topography in 1992 and operated until 2006. Further satellites have been launched in follow-on missions to continue monitoring the global sea surface: Jason-1 in 2001, Jason-2 in 2008, and Jason-3 in 2016. The satellites use radar altimetry to measure the height of the satellite above the ocean surface. Satellite laser ranging from 10–15 ground-based ranging stations and on-board GPS locate the position of the satellite very accurately, so that the height of the reference geoid beneath the satellite can be determined. The difference in heights (up to a meter, with precision of a few millimeters) defines the highs and lows of the ocean surface topography. Gravity causes water to flow away from a topographic high and toward a low, so the satellite data allow the motion of oceanic currents to be mapped, providing important information about ocean circulation and associated climatic effects.

The GRACE satellites repeatedly measure the mass of the ice trapped in glaciers and in the great ice-sheets covering Greenland and Antarctica. The weight of the ice causes an isostatic response in the underlying crust and upper mantle (Section 4.3). This introduces uncertainty into the calculation of the mass of ice in each of the ice-sheets, but a correction is possible. The isostatic correction does not change during the short duration of satellite observations so it does not affect seasonal differences. GRACE measurements over several years from 2002 to 2009 show seasonal fluctuations as ice-mass is gained in winter and lost in summer. The fluctuations are superposed on a long-term trend that indicates progressive loss of mass in both ice-sheets during the measurement interval (Fig. 3.33). The mean annual loss of ice-mass is estimated to be around 140–230 Gt yr^{-1} (one gigaton, Gt, equals 10^9 tons) in Greenland and 70–140 Gt yr^{-1} in Antarctica. The ice-mass differences in Antarctica are complicated by the observation that East Antarctica is gaining ice-mass while West Antarctica is losing it at a faster rate, so that a net annual loss ensues. The melted ice from both ice-sheets is responsible for a large fraction of the global increase in sea level, amounting to 0.5–1.0 mm yr^{-1}. In both Greenland and Antarctica the long-term loss in ice-mass is better represented by a quadratic function than by a linear decrease. This indicates that the ice loss due to melting of the ice-sheets is accelerating.

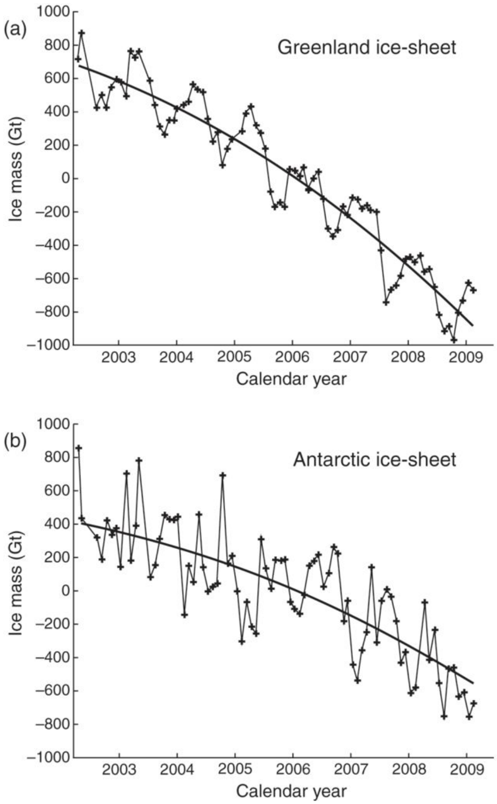


Fig. 3.33 Mass loss due to melting of the ice-sheets that cover (a) Greenland and (b) the Antarctic. The smooth curve is a best-fitting quadratic, which shows that the mass loss is accelerating for each ice-sheet.

Source: I. Velicogna, Increasing rates of ice mass loss from the Greenland and Antarctic ice sheets revealed by GRACE, *Geophys. Res. Lett.*, L19503, 2009.

Satellite records of the gravitational field also document environmentally important changes in mass distribution caused by human activity. For example, repeated GRACE measurements over major aquifers reveal short-term loss of mass due to the depletion of groundwater by pumping for irrigation. This is evident in GRACE results for California’s Central Valley, northeastern India, central Australia, and the Middle East. The results are important for planning the sustainable use of water resources.

3.5.7 Observations of Crustal Deformation with Synthetic Aperture Radar

Among the many satellites in Earth orbit, some (identified by acronyms such as ERS1, ERS2, JERS, IRS, RADARSAT, Envisat, Sentinel, etc.) carry equipment specifically designed to direct beams of radar waves at the Earth and record the reflections from the Earth’s surface. Synthetic aperture radar (SAR) is a remote-sensing technique that has made it possible to record features of the Earth’s surface in remarkable detail based on these radar reflections.

Radar signals, like visible light, are subject to reflection, refraction, and diffraction (these phenomena are explained in Section 6.3.2 for seismic waves). Diffraction (see Fig. 6.18) bends light passing through a lens in such a way that a point source becomes diffuse. When two adjacent point sources are observed with the lens, their diffuse images overlap, and if they are very close they may not be seen as distinct points. The resolving power of an optical instrument, such as a lens, is defined by the smallest angular separation (θ) of two points that the instrument can distinguish clearly. For a given lens this angle is dependent inversely on the diameter (d) of the aperture that allows light to pass through the lens, and directly on the wavelength (λ) of the light. It is given by the approximate relationship $\theta \sim \lambda/d$. High resolution requires that closely spaced details of the object can be distinguished, i.e., the angular resolution θ should be a small number. Thus, the larger the aperture of the lens, the higher is the optical resolution.

The same principle applies to radar. Instead of being dependent on the diameter of an optical lens, the resolution of a radar system is determined by the antenna length. When mounted in a satellite, the physical dimensions of the antenna are limited to just a few meters. SAR makes use of the motion of the antenna and powerful data processing to get around this limitation.

The radar antenna is mounted so that it directs its beam at right angles to the direction of motion of the host spacecraft. The beam “illuminates” a swathe of ground surface, each particle of which reflects a signal to the antenna. Hundreds of radar pulses are sent out per second (e.g., the European Radar Satellites [ERS] emit 1700 pulses per second); this results in huge amounts of reflected signals. As the craft moves forward, the illuminated swathe moves across the target surface. Each particle of the target reflects hundreds of radar pulses from the time when it is first energized until it is no longer covered by the beam. During this time the craft (and real antenna) move some distance along the track. In subsequent data processing the signals reflected from the target are combined and corrected for the changing position of the antenna in such a way that they appear to have been gathered by an antenna as long as the distance moved along the track. This distance is called the *synthetic aperture* of the radar. For example, a SAR investigation with an ERS1 satellite in orbit 800 km above the Earth’s surface created a synthetic aperture of about 4 km. The high resolving power achieved with this large aperture produced SAR images of ground features with a resolution of about 30 m.

An important aspect of the data reduction is the ability to reconstruct the path of each reflection precisely. This is achieved by using the Doppler effect, the principle of which is described in Box 1.1. Reflections from target features ahead of the moving spacecraft have elevated frequencies; those from behind have lowered frequencies. Correcting the frequency of each signal for its Doppler shift is necessary to obtain the true geometry of the reflections.

A further development of the SAR method is *interferometric* SAR (InSAR). This technique analyzes the phase of the reflected radar signal to determine small changes in topography between repeated passages of the satellite over an area. The phase of the wave is a measure of the time delay the wave experiences in transit between transmitter and receiver. To illustrate this point, picture the shape of a waveform as an alternation of crests and troughs, which leaves the satellite at the instant its amplitude A is maximum (i.e., at a crest). If the reflected signal returns to the satellite as a crest, it has the same phase as the transmitted signal. Its amplitude at time t can be expressed by the equation $y = A\cos\omega t$, where ω is the signal frequency. This will be the case if the path of the reflected wave to and from the target is an exact number of complete wavelengths. On the other hand, if the reflection arrives back at the satellite as a trough, it is exactly out-of-phase with the original wave. This happens when the length of its path is an odd number of half-wavelengths. More generally, the path length is not an exact (even or odd) number of half-wavelengths, and the equation of the reflected wave must be written $y = A\cos(\omega t + \Delta)$, where the phase difference Δ depends on the path length. The InSAR technique developed in the 1990s is based on analysis of the phase differences inherent in each reflection recorded by the satellite.

If a SAR image is made of a target area during one orbit, it should be reproduced exactly on a later orbit that revisits the same location (this is not exactly possible, but paths that repeat within a few hundred meters can be corrected for geometric differences). In particular, because each point of the target is the same distance from the transmitter, the phases of the imaged signals should be identical. However, if geological (or man-made) events have caused surface displacements between the times of the two images there will be phase differences between them. These are made visible by combining the two images so that they interfere with each other.

When harmonic signals with different phases are mixed, they interfere with each other. Constructive interference occurs when the signals have the same phase; if they are superposed, the combined signal is strengthened. Destructive interference occurs when out-of-phase signals are mixed; the combined signal is weakened. The interference pattern that results from mixing the two waveforms consists of alternating zones of reinforcement and reduction of the signal, forming a sequence of so-called “interference fringes.”

When this procedure is carried out with SAR images, the resulting interference pattern makes it possible to interpret ground motion of large areas in much greater detail than would be possible from ground-based observations. The method has been used to record various large-scale ground displacements related, for example, to earthquakes, tectonic faulting, and volcanism. Fig. 3.34 shows an interference pattern superposed on the background topography of Mount Etna, in Sicily, following a cycle of eruptions in 1992 and 1993. Successive radar images from the same point were obtained 13 months apart by the ERS1 satellite, which transmitted radar signals with wavelength 5.66 cm. In order to change the line-of-sight distance between the satellite and the target by a full wavelength, the ground must move by a half-wavelength along this direction, in this case by 2.83 cm. The concentric dark and light fringes around the crater show four cycles of interference, corresponding to an elevation change of the mountaintop by about 11 cm. The fringes result from the subsidence of the crater as magma drained out of it following the eruptive cycle.

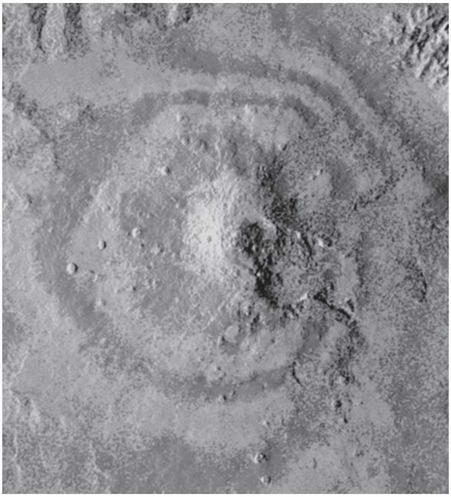


Fig. 3.34 Interferometric synthetic aperture radar (InSAR) pattern of interference fringes showing changes in elevation of Mount Etna, Sicily, following the 1992–1993 eruptive cycle. The four pairs of light and dark fringes correspond to subsidence of the mountaintop by about 11 cm as magma drained out of the volcano.

Source: D. Massonnet, Satellite radar interferometry, *Scientific American* 276(2), 46–53, 1997.

3.5.8 Very Long Baseline Interferometry

Extra-galactic radio sources (quasars) form the most stable inertial coordinate system yet known for geodetic measurements. The extra-galactic radio signals are detected almost simultaneously by radio-astronomy antennas at observatories on different continents. Knowing the direction of the incoming signal, the small differences in times of arrival of the signal wavefronts at the various stations are processed to give the lengths of the baselines between pairs of stations. This highly precise geodetic technique, called very long baseline interferometry (VLBI), allows determination of the separation of observatories several thousand kilometers apart with an accuracy of a few centimeters. Although not strictly a satellite-based technique, it is included in this section because of its use of non-terrestrial signals for high-resolution geodetic measurements.

By combining VLBI observations from different stations, the orientation of the Earth to the extra-galactic inertial coordinate system of the radio sources is obtained. Repeated determinations yield a record of the Earth's orientation and rotational rate with unprecedented accuracy. Motion of the rotation axis (e.g., the Chandler wobble, Section 3.3.5.3) can be described optically with a resolution of 0.5–1 m; the VLBI data have an accuracy of 3–5 cm. The period of angular rotation can be determined to better than 0.1 ms. This has enabled very accurate observation of irregularities in the rotational rate of the Earth, which are manifest as changes in the length of the day (LOD).

The most important, first-order changes in the LOD are due to the braking of the Earth's rotation by the lunar and solar marine tides (Section 3.3.4). The most significant non-tidal LOD variations are associated with changes in the angular momentum of the atmosphere due to shifts in the east–west component of the wind patterns. To conserve the total angular momentum of the Earth, a change in the angular momentum of the atmosphere must be compensated by an equal and opposite change in the angular momentum of the crust and mantle. The largely seasonal transfers of angular momentum correlate well with high-frequency variations in the LOD obtained from VLBI results (Fig. 3.35).

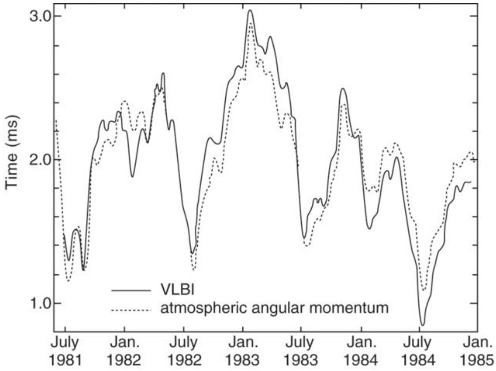


Fig. 3.35 Fine-scale fluctuations in the LOD observed by VLBI, and LOD variations expected from changes in the angular momentum of the atmosphere.

Source: W.E. Carter, Earth orientation, in: *The Encyclopedia of Solid Earth Geophysics*, D.E. James, ed., pp. 231–239, New York: Van Nostrand Reinhold, 1989.

If the effects of marine tidal braking and non-tidal transfers of atmospheric angular momentum variations are taken into account, small residual deviations in the LOD remain. These are related to the tides in the solid Earth (Section 3.3.4.5). The lunar and solar tidal forces deform the Earth elastically and change its ellipticity slightly. The readjustment of the mass distribution necessitates a corresponding change in the Earth’s rate of rotation in order to conserve angular momentum. The expected changes in LOD due to the influence of tides in the solid Earth can be computed. The discrepancies in LOD values determined from VLBI results (Fig. 3.36) agree well with the fluctuations predicted by the theory of the bodily Earth tides.

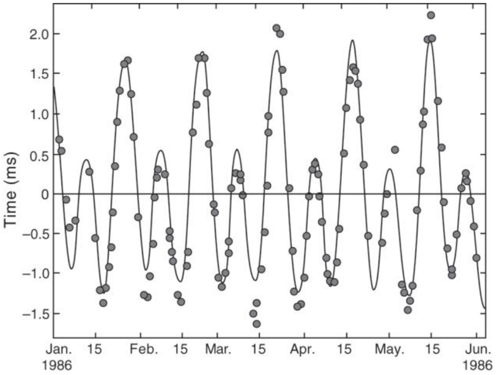


Fig. 3.36 High-frequency changes in the LOD after correction for the effects due to atmospheric angular momentum (points) and the theoretical variations expected from the solid bodily Earth tides.

Source: W.E. Carter, Earth orientation, in: *The Encyclopedia of Solid Earth Geophysics*, D.E. James, ed., pp. 231–239, New York: Van Nostrand Reinhold, 1989.

3.6 *Suggestions for Further Reading*

Introductory Level

Grotzinger, J. and Jordan, T. H. 2014. *Understanding Earth* (7th ed.). San Francisco, CA: W.H. Freeman.

Massonnet, D. 1997. Satellite radar interferometry. *Scientific American*, **276**, 46–53.

Intermediate and Advanced Level

Bullen, K. E. 1975. *The Earth's Density*. London: Chapman and Hall.

Fowler, C. M. R. 2004. *The Solid Earth: An Introduction to Global Geophysics* (2nd ed.). Cambridge: Cambridge University Press.

Stacey, F. D. and Davis, P. M. 2008. *Physics of the Earth* (4th ed.). Cambridge: Cambridge University Press.

Turcotte, D. L. and Schubert, G. 2014. *Geodynamics* (3rd ed.). Cambridge: Cambridge University Press.

3.7 Review Questions

1. What is the *geoid*? What is the *reference ellipsoid*? How and why do they differ?

2. What is a geoid anomaly? Explain how a positive (or negative) anomaly arises.

3. What is *normal gravity*? What does the word “normal” imply? Which surface is involved?

4. Gravitational acceleration is directed toward a center of mass. With the aid of a sketch that shows the directions of gravity’s components, explain why *gravity* is not a centrally directed acceleration.

5. Explain which geophysical parameters determine each of the constants in the normal gravity formula?

6. What is the Coriolis acceleration? How does it originate? How does it affect wind patterns in the north and south hemispheres?

7. What is the Eötvös acceleration? How does it originate? How does it affect a gravity measurement?

8. Explain how to calculate the position of the common center of mass of the Earth–Moon system.

9. Explain with the aid of diagrams that show the forces involved why there are two lunar tides per day.

10. Why are the lunar tides almost equal on opposite sides of the Earth? Why are they not exactly equal?

11. What are bodily Earth tides, and how are they measured?

12. What causes the precession of the Earth’s rotation axis? Why is it retrograde?

13. What other long-term changes of the rotation axis or the Earth’s orbit occur? What are the periods of these motions? What are their causes?

14. What are the roles of laser ranging and satellite altimetry in determining the dynamic topography of the ocean surface?

15. How is gravity measured in the GRACE and GOCE satellite missions?

3.8 Exercises

1. Using the data in Table 1.1, calculate the gravitational acceleration on the surface of the Moon as a percentage of that on the surface of the Earth.

2. An Olympic high-jump champion jumps a record height of 2.45 m on the Earth. How high could this champion jump on the Moon?

3. (a) Calculate the velocity a rocket must have in order to escape from the Earth's gravity, assuming a mean gravitational acceleration of 9.81 m s^{-1} and mean Earth radius of 6371 km.

(b) What is the escape velocity of the same object on the Moon?

(c) What velocity would a rocket fired from the Earth need to have in order to escape from the solar system?

4. The equatorial radius of the Earth is 6378 km and gravity at the equator is 9.780 m s^{-2}. Compute the ratio μ of the centrifugal acceleration at the equator to the *gravitational acceleration* at the equator. If the ratio μ is written as $1/k$, what is the value of k?

5. Given that the length of a sidereal month is 27.32 solar days, the mean gravity on Earth is 9.81 m s^{-2} and the Earth's radius is 6371 km, calculate the radius of the Moon's orbit.

6. A communications satellite is to be placed in a geostationary orbit.

(a) What must the period and orientation of the orbit be?

(b) What is the radius of the orbit?

(c) If a radio signal is sent to the satellite from a transmitter at latitude 45°N, what is the shortest time taken for its reflection to reach the Earth?

7. Calculate the centrifugal acceleration due to the Earth's rotation of an object at rest on the Earth's surface in Paris, assuming a latitude of 48° 52′ N. Express the result as a percentage of the gravitational attraction on the object.

8. A solid angle (Ω) is defined as the quotient of the area (A) of the part of a spherical surface subtended by the angle, divided by the square of the spherical radius (r): i.e., $\Omega = A/r^2$ (see Box 11.4). Show with the aid of a diagram that the gravitational acceleration at any point surrounded by a thin, homogeneous, spherical shell is zero.

9. Assuming that the gravitational acceleration inside a homogeneous spherical shell is zero, show that the gravitational acceleration inside a homogeneous uniform *solid* sphere is proportional to the distance from its center.

10. Show that the gravitational potential U_G inside a homogeneous uniform solid sphere of radius R at a distance r from its center is given by

$$U_G = -\frac{2}{3}G\rho(3R^2 - r^2)$$

11. Sketch the variations of gravitational acceleration and potential inside and outside a homogeneous solid sphere of radius R.

12. A thin borehole is drilled through the center of the Earth, and a ball is dropped into the borehole. Assume the Earth to be a homogeneous solid sphere. Show that the ball will oscillate back and forth from one side of the Earth to the other. How long does it take to traverse the Earth and reach the other side?

13. The Roche limit is the closest distance an object can approach a planet before being torn apart by the tidal attraction of the planet. For a rigid spherical moon the Roche limit is given by Eq.

(6) in Box 3.2.

(a) Using the planetary dimensions in Table 1.1, calculate the Roche limit for the Moon with respect to the Earth. Express the answer as a multiple of the Earth's radius.

(b) Show that, for a planet whose mean density is less than half that of its rigid moon, the moon would collide with the planet before being torn apart by its gravity.

(c) Given that the Sun's mass is 1.989×10^{30} kg and that its radius is 695,500 km, calculate the Roche limit for the Earth with respect to the Sun.

(d) The mean density of a comet is about 500 kg m^{-3}. What is the Roche limit for comets that might collide with the Earth?

(e) The mean density of an asteroid is about 2000 kg m^{-3}. If an asteroid on a collision course with the Earth has a velocity of 15 km s^{-1}, how much time will elapse between the break-up of the asteroid at the Roche limit and the impact of the remnant pieces on the Earth's surface, assuming they maintain the same velocity as the asteroid?

14. The mass *M* and moment of inertia *C* of a thick shell of uniform density ρ, with internal radius *r* and external radius *R* are given by

$$M = \frac{4}{3}\pi\rho(R^3 - r^3) \qquad C = \frac{8}{15}\pi\rho(R^5 - r^5)$$

The Earth has an internal structure consisting of concentric spherical shells. A simple model with uniform density in each shell is given in the following figure:

Layer	Radius (km)	Density (kg m^{-3})
	6370	
upper mantle		3300
	5700	
lower mantle		5000
	3480	
outer core		11000
	1220	
inner core		13000
	0	

(a) Compute the mass and moment of inertia of each spherical shell.

(b) Compute the total mass and total moment of inertia of the Earth.

(c) If the moment of inertia can be written $C = kMR^2$, where M is Earth's mass and R its radius, what is the value of k?

(d) What would be the value of k if the density were uniform throughout the Earth?

15. The gravity potential for a rotating sphere (or spheroid) is given by Eq. (3.56). Show that this expression does not satisfy the Laplace equation, and give reasons why this is the case. (Hint: Use spherical coordinates as in Box 3.4, Eq. (2), but assume azimuthal symmetry about the rotation axis, i.e., the potential does not vary with ϕ.)

3.9 Computational Exercises

Jupyter notebook computational exercises can be found online for this chapter at www.cambridge.org/FoG3ed ((1) Gravity of an extended body, (2) Tides and Roche limit).

4

Gravity Surveying

Preview

Gravity measurements are made with a gravimeter at geodetically located places on land or on gyroscopically stabilized platforms in marine and airborne surveys. In airborne and satellite surveys, gravity gradiometers are used. We explain these instruments and how gravity data are reduced to the reference ellipsoid by correcting for the latitude and altitude of the measuring site and the attraction of surrounding topography. A residual gravity anomaly is due to subsurface structure. We distinguish two types of anomaly: free-air anomalies are not corrected for local rock densities, which are needed to define Bouguer anomalies. The interpretation of local anomalies is facilitated by modeling the gravitational attractions of simple geometric bodies. Regional gravity anomalies help us to understand the geological processes at mountain chains and plate boundaries. Isostasy is an important dynamic process, by which the excess mass of a mountain is compensated by a less dense subsurface structure. Vertical crustal movements occur when isostatic equilibrium is disturbed.

4.1 Gravity Measurement and Reduction

4.1.1 Absolute and Relative Measurement of Gravity

The mean value of gravity at the surface of the Earth is approximately 9.81 m s^{-2}, or 981,000 mGal. The Earth's rotation and flattening cause gravity to increase by roughly 5300 mGal from equator to pole, which is a variation of only about 0.5%. Accordingly, measurements of gravity are of two types. The first corresponds to determination of the *absolute* magnitude of gravity at any place; the second consists of measuring the *change* in gravity from one place to another. In geophysical studies, especially in gravity prospecting, it is necessary to measure accurately the small changes in gravity caused by underground structures. These require an instrumental sensitivity of the order of 0.01 mGal. It is very difficult to design an instrument to measure the absolute value of gravity that has this high precision and that is also portable enough to be used easily in different places. Gravity surveying is usually carried out with a portable instrument called a *gravimeter*, which determines the variation of gravity relative to one or more reference locations. In national gravity surveys the relative variations determined with a gravimeter may be converted to absolute values by calibration with absolute measurements made at selected stations.

4.1.2 Absolute Measurement of Gravity

The classical method of measuring gravity is with a pendulum. A simple pendulum consists of a heavy weight suspended at the end of a thin fiber. The compound (or reversible) pendulum, first described by Henry Kater in 1818, allows more exact measurements. It consists of a stiff metal or quartz rod, about 50 cm long, to which is attached a movable mass. Near each end of the rod is fixed a pivot, which consists of a quartz knife-edge resting on a flat quartz plane. The period of the pendulum is measured for oscillations about one of the pivots. The pendulum is then inverted and its period about the other pivot is determined. The position of the movable mass is adjusted until the periods about the two pivots are equal. The distance *L* between the pivots is then measured accurately. The period of the instrument is given by

$$T = 2\pi\sqrt{\frac{I}{mgh}} = 2\pi\sqrt{\frac{L}{g}}\,, \tag{4.1}$$

where *I* is the moment of inertia of the pendulum about a pivot, *h* is the distance of the center of mass from the pivot, and *m* is the mass of the pendulum. Knowing the length *L* from Kater's method obviates knowledge of *I, m*, and *h.*

The sensitivity of the compound pendulum is found by differentiating Eq. (4.1). This gives

$$\frac{\Delta g}{g} = -2\frac{\Delta T}{T}\,. \tag{4.2}$$

To obtain a sensitivity of about 1 mGal it is necessary to determine the period with an accuracy of about 0.5 μs. This can be achieved easily today with precise atomic clocks. The compound pendulum was the main instrument for gravity prospecting in the 1930s, when timing the swings precisely was more difficult. It was necessary to time as accurately as possible a very large number of swings. As a result a single gravity measurement took about half an hour.

The performance of the instrument was handicapped by several factors. The inertial reaction of the housing to the swinging mass of the pendulum was compensated by mounting two pendulums on the same frame and swinging them in opposite phase. Air resistance was reduced by housing the pendulum assemblage in an evacuated thermostatically controlled chamber. Friction in the pivot was minimized by the quartz knife-edge and plane, but due to minor unevenness the contact edge was not exactly repeatable if the assemblage was set up in a different location, which affected the reliability of the measurements. The apparatus was bulky but was used until the 1950s as the main method of making absolute gravity measurements.

4.1.2.1 Free-Fall Method

Modern methods of determining the acceleration of gravity are based on observations of falling objects. For an object that falls from a starting position z_0 with initial velocity u, the equation of motion gives the position z at time t as

$$z = z_0 + ut + \frac{1}{2}gt^2 . \tag{4.3}$$

The absolute value of gravity is obtained by fitting a quadratic to the record of position versus time.

An important element in modern experiments is the accurate measurement of the change of position with a Michelson interferometer. In this device a beam of monochromatic light passes through a beam splitter, consisting of a semi-silvered mirror, which reflects half of the light incident upon it and transmits the other half. This divides the incident ray into two subrays, which subsequently travel along different paths and are then recombined to give an interference pattern. If the path lengths differ by a full wavelength (or a number of full wavelengths) of the monochromatic light, the interference is constructive. The recombined light has maximum intensity, giving a bright interference fringe. If the path lengths differ by half a wavelength (or by an odd number of half- wavelengths) the recombined beams interfere destructively, giving a dark fringe. In modern experiments the monochromatic light source is a laser beam of accurately known wavelength.

In an absolute measurement of gravity a laser beam is split along two paths that form a Michelson interferometer (Fig. 4.1). The horizontal path is of fixed length, while the vertical path is reflected off a corner-cube retroreflector that is released at a known instant and falls freely. The path of free-fall is about 0.5 m long. The cube falls in an evacuated chamber to minimize air resistance. A photomultiplier and counter permit the fringes within any time interval to be recorded and counted. The intensity of the recombined light fluctuates sinusoidally with increasing frequency the further and faster the cube falls. The distance between each zero crossing corresponds to half the wavelength of the laser light, and so the distance traveled by the falling cube in any time interval may be obtained. The times of the zero crossings must be measured with an accuracy of 0.1 ns (10^{-10} s) to give an accuracy of 1 μGal in the gravity measurement.

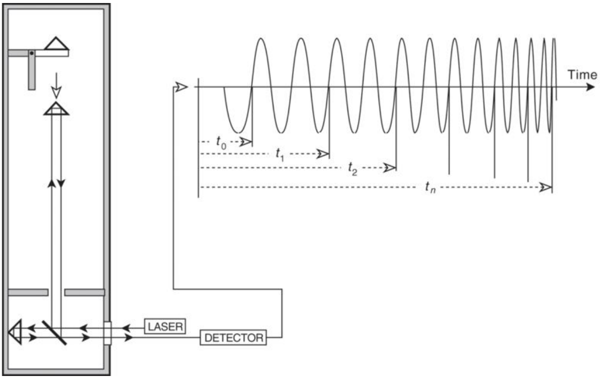


Fig. 4.1 The free-fall method of measuring absolute gravity.

Although the apparatus is compact, it is not quite portable enough for gravity surveying. It gives measurements of the absolute value of gravity with an accuracy of about 0.005–0.010 mGal (5–10 μGal). A disadvantage of the free-fall method is the resistance of the residual air molecules left in the evacuated chamber. This effect is reduced in the rise-and-fall method.

4.1.2.2 Rise-and-Fall Method

In the original version of the rise-and-fall method a glass sphere was fired vertically upward and fell back along the same path (Fig. 4.2). Timing devices at two different levels registered the times of passage of the ball on the upward and downward paths. In each timer a light-beam passed through a narrow slit. As the glass sphere passed the slit it acted as a lens and focused one slit on the other. A photomultiplier and detector registered the exact passage of the ball past the timing level on the upward and downward paths. The distance h between the two timing levels (around 1 m) was measured accurately by optical interferometry.

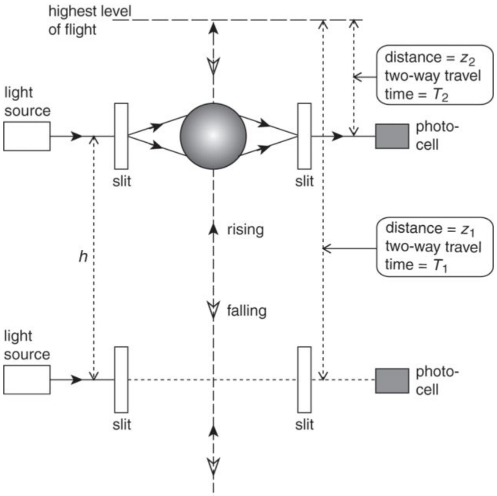

highest level of flight
light source
slit
slit
photo-cell
distance = z_2 two-way travel time = T_2
rising
falling
h
distance = z_1 two-way travel time = T_1
light source
slit
slit
photo-cell

Fig. 4.2 The rise-and-fall method of measuring absolute gravity.

Let the time spent by the sphere above the first timing level be T_1 and the time above the second level be T_2; further, let the distances from the zenith level to the timing levels be z_1 and z_2, respectively. The corresponding times of fall are $t_1 = T_1/2$ and $t_2 = T_2/2$. Then,

$$z_1 = \frac{1}{2} g \left(\frac{T_1}{2} \right)^2, \tag{4.4}$$

with a similar expression for the second timing level. Their separation is

$$h = z_1 - z_2 = \frac{1}{8} g (T_1^2 - T_2^2). \tag{4.5}$$

The following elegantly simple expression for the value of gravity is obtained:

$$g = \frac{8h}{(T_1^2 - T_2^2)}. \tag{4.6}$$

Although the experiment is conducted in a high vacuum, the few remaining air molecules cause a drag that opposes the motion. On the upward path the air drag is downward, in the same direction as gravity; on the downward path the air drag is upward, opposite to the direction of gravity. This asymmetry helps to minimize the effects of air resistance.

In a modern variation, Michelson interferometry is used as in the free-fall method. The projectile is a corner-cube retroreflector (see Box 1.4), and interference fringes are observed and counted during its upward and downward paths. Sensitivity and accuracy are comparable to the free-fall method.

4.1.2.3 Absolute Gravity Networks

Early in the twentieth century comparisons between absolute gravity measurements at different places were made with the aid of pendulums. They showed discrepancies of many mGals, which became obvious when accurate spring-based gravimeters became operational. The development of the free-fall absolute gravimeter made a worldwide network possible, consisting of more than 100 stations, in which absolute gravity measurements are made under standardized conditions. The worldwide accuracy with which absolute measurements can be made is around 5 μGal.

4.1.3 Relative Measurement of Gravity: The Gravimeter

In principle, a gravity meter, or *gravimeter*, is a very sensitive balance. The first gravimeters were based on the straightforward application of Hooke's law (Section 5.1.1). A mass m suspended from a spring of length s_0 causes it to stretch to a new length s. The extension, or change in length, of the spring is proportional to the restoring force of the spring and so to the value of gravity, according to

$$F = mg = -k(s - s_0), \tag{4.7}$$

where k is the elastic constant of the spring. The gravimeter is calibrated at a known location. If gravity is different at another location, the extension of the spring changes, and from this the change in gravity can be computed.

This type of gravimeter, based directly on Hooke's law, is called a stable type. It has been replaced by more sensitive unstable or astatized types, which are constructed so that an additional force acts in the same direction as gravity and opposes the restoring force of the spring. The instrument is then in a state of unstable equilibrium. This condition is realized through the design of the spring. If the natural length s_0 can be made as small as possible, ideally zero, Eq. (4.7) shows that the restoring force is then proportional to the physical length of the spring instead of its extension. The *zero-length spring,* first introduced in the LaCoste–Romberg gravimeter, is now a common element in modern gravimeters. The spring is usually of the helical type. When a helical spring is stretched, the fiber of the spring is twisted; the total twist along the length of the fiber equals the extension of the spring as a whole. During manufacture of a zero-length spring the helical spring is given an extra twist, so that its tendency is to uncoil. An increase in gravity stretches the spring against its restoring force, and the extension is augmented by the built-in pre-tension.

The operation of a gravimeter is illustrated in Fig. 4.3. A mass is supported by a horizontal rod to which a mirror is attached. The position of the rod is observed with a light-beam reflected into a microscope. If gravity changes, the zero-length spring is extended or shortened and the position of the rod is altered, which deflects the light-beam. The null-deflection principle is utilized. An adjusting screw changes the position of the upper attachment of the spring, which alters its tension and restores the rod to its original horizontal position as detected by the light-beam and microscope. The turns of the adjusting screw are calibrated in units of the change in gravity, usually in mGal.

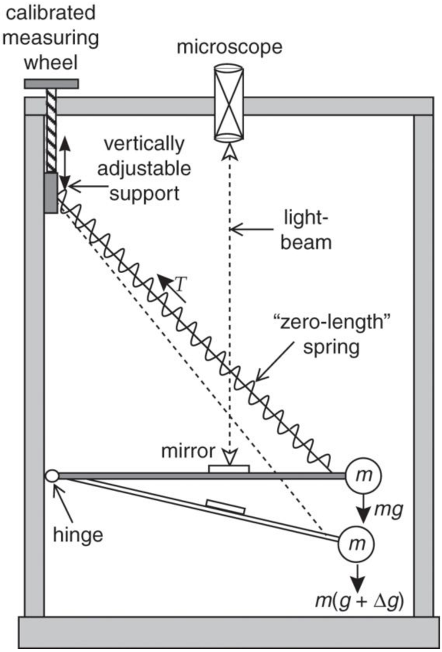

calibrated measuring wheel
microscope
vertically adjustable support
light-beam
T
"zero-length" spring
mirror
m
mg
hinge
m
m(g + Δg)

Fig. 4.3 The principle of operation of an unstable (astatic) type of gravimeter.

The gravimeter is light, robust, and portable. After initially leveling the instrument, an accurate measurement of a gravity difference can be made in a few minutes. The gravimeter has a sensitivity of about 0.01 mGal (10 μgal). This high sensitivity makes it susceptible to small changes in its own properties.

4.1.3.1 The Superconducting Gravimeter

During the past half-century the superconducting gravimeter has been developed, utilizing the special characteristics of superconductivity (Box 4.1). Its sensor consists of a sphere of superconducting niobium suspended in the magnetic field gradient created by a pair of superconducting coils. A constant current is trapped in the coils, producing a persistent field. A field gradient is obtained by using the geometry of the coils (e.g., by placing the sphere just outside the space between them) or by using unequal coils. The gradient produces a force on the sphere, which holds it in position. If gravity changes, the position of the sphere moves relative to the coils; this is detected by an electrical capacitance bridge in which one arm is the coil. The imbalance of the bridge activates a feedback current so that the sphere is kept suspended in position. The force to maintain this levitation is calibrated accurately in terms of the change in gravity.

Box 4.1 Superconductivity

As the temperature of a normal material is lowered, the collisions between the free electrons and the atoms in the lattice of the material, which are responsible for its electrical resistance, decrease. In a superconducting material the electrons are not free but are paired with a partner so that they share the same quantum state, which is not possible for individual electrons. The pairing is enabled by an adjustment of the lattice vibrations (*phonons*) of the material. Below a critical temperature there is insufficient thermal energy to scatter the “pair,” which can then move unopposed through the material like a single charged particle. The electrical resistance of the material is zero below the critical temperature and it is superconducting.

An electrical current can persist indefinitely in a superconductor, which results in many uses. A strong current in a superconducting coil can be used to produce very strong magnetic fields. This feature has many practical applications, from magnetic resonance imaging (MRI) in medical care to the development of high-speed magnetic-levitation trains. Very *strong* magnetic fields created in superconducting electromagnets are a crucial element in the design of high-energy particle accelerators, such as the Large Hadron Collider. On the other hand, superconductivity is the fundamental property in the design of the Superconducting QUantum Interference Device (SQUID), which is used for detecting and measuring very *weak* magnetic fields. The infinite conductivity of a superconductor is also an important factor in magnetic shielding, which creates a space that is entirely free from magnetic fields. This is possible because an external magnetic field cannot enter a superconductor.

The principle of operation is similar to the LaCoste–Romberg gravimeter, but with the spring replaced by a levitating sphere. In contrast to a normal gravimeter, which is subject to instrumental drift (Section 4.1.4), the superconducting gravimeter is virtually drift-free. However, the need for thermal insulation and a supply of liquid helium make the instrument more suitable for installation in an observatory than for fieldwork. It is well suited for the long-term measurement of gravity at a fixed location because its noise level is around 0.1 μGal in short-term analyses and in the range of 0.01 μGal for long-term observations. It can be used for observations of bodily Earth tides and the Chandler wobble (Sections 3.3.4 and 3.3.5), as well as the normal modes of the Earth’s natural oscillations (Section 6.2.4).

4.1.3.2 The Gravity Gradiometer

The inverse square law causes gravitational attraction to decrease with distance from the Earth's center. This causes gravity to decrease in the vertical direction. The rate of decrease is the *vertical* gravity gradient; it amounts to about 0.3086 mGal per meter of elevation (Section 4.1.5.5). Gravity acts normal to the ellipsoid and not radially, so there are also *horizontal* gravity gradients. The normal gravity formula (Section 3.4.4) allows the change of gravity with latitude to be calculated by differentiating the formula. This horizontal gradient of gravity amounts to about 0.008 mGal per meter of north–south displacement. Other gravitational attractions, arising from topography or underground inhomogeneities, create vertical and horizontal gravity gradients. The unit of measurement of gravity gradient is the eotvos (named after Loránd Eötvös, the Hungarian developer of the method in the late nineteenth century). The eotvos (symbol E) is defined as an acceleration gradient of 1 nanoGal/cm (dimensionally equivalent to s^{-2}). The vertical gradient of gravity at Earth's surface is thus equivalent to 3086 E.

Gravity gradients are measured directly with a gradiometer. Gravity gradiometry is an important method used in exploring for mineral resources, as well as being a key technique in global analyses of the gravitational field from spacecraft. The sensors in a gradiometer are sensitive accelerometers (Box 4.2) employed in pairs with a common axis of measurement. A gradiometer is designed for gravity measurements from a moving vehicle. Both sensors react together in the case of a disturbance, such as turbulence in an aircraft or stormy conditions in a marine investigation. Because they move in unison the difference between the signals from the two accelerometers is small or zero, and in a uniform gravitational field both accelerometers sense the same field. However, if one element of the pair is closer than the other to a disturbing mass, it experiences a stronger attraction. The difference between the accelerometer signals divided by their separation is the gravity gradient along the axis of the accelerometer pair.

Box 4.2 The Accelerometer

The most commonly used types of accelerometer are the capacitive and electrostatic type. The structure of the capacitive type (Fig. B4.2.1a) helps explain the principle of operation. A proof mass is suspended between the plates of a capacitor, with a small distance d between its upper surface and the top plate. If the device is stationary or moving at constant velocity, the gaps between the proof mass and the capacitor plates are constant. A downward acceleration, e.g., an increase in gravity, causes the proof mass to move downward, increasing the separation d_1; this reduces the capacitance C_1 between the upper plate and the top of the mass. At the same time the gap d_2 between the bottom of the mass and the lower plate decreases, increasing the corresponding capacitance C_2. The capacitance of the assemblage is changed by the imbalance and is corrected by a servo-controlled current so that the mass is held in a neutral position. The restorative current is a measure of the gravity change that caused the imbalance.

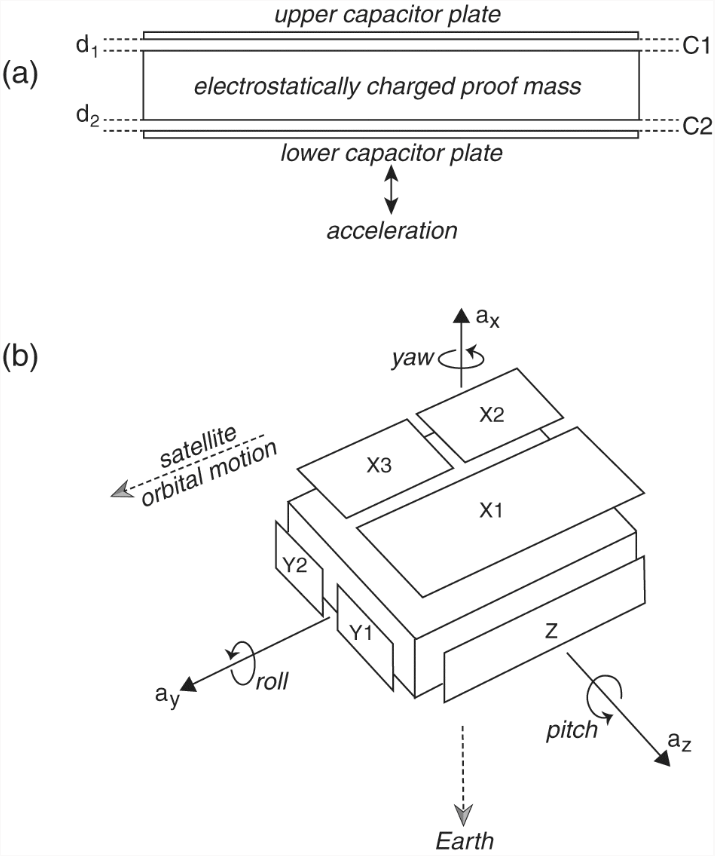

(a)
upper capacitor plate
d1
C1
electrostatically charged proof mass
d2
C2
lower capacitor plate
acceleration
(b)
ax
yaw
X2
X3
X1
satellite orbital motion
Y2
Y1
Z
ay
roll
pitch
az
Earth

Fig. B4.2.1 (a) Principle of operation of an electrostatic accelerometer. (b) Sketch of the arrangement of six pairs of electrodes (X1, X2, X3, Y1, Y2, and Z) surrounding a levitating proof mass in the STAR accelerometer on the CHAMP spacecraft (after Perosanz et al., 2003).

The accelerometer can be constructed more robustly with piezoelectric crystals in place of the capacitor plates to detect the motion of the test mass. A change in gravity causes a change in the pressure of the test mass on the piezoelectric crystals, producing an electrical current. A feedback system again holds the test mass in place and gives a measure of the change in gravity. The accelerometer in this design has no moving parts and can be used for a multitude of purposes, from everyday use as a motion sensor in mobile phones to scientific operations on board spacecraft.

An accelerometer responds to acceleration in a known direction. A three-dimensional (3-D) accelerometer involves the simultaneous operation of three orthogonal axial accelerometers. The STAR (***S****pace* ***T****hree-axis* ***A****ccelerometer for* ***R****esearch*) accelerometer on the CHAMP satellite was located at the center of mass of the satellite. Its electrically charged proof mass consisted of a rectangular block made of a titanium alloy, which levitated electrostatically between three orthogonal pairs of capacitor plates that measured directional accelerations (Fig. B4.2.1b). Six pairs of electrodes detect the motion of the electrostatically levitated proof mass. The electrodes had different sizes and were arranged off-center so that rotational accelerations about the orthogonal axes could also be measured. Combinations of the voltages gave the accelerations along the sensor axes and the rotational accelerations (pitch, roll, and yaw) about these axes, so that the gravity measurements could be corrected for accelerations of the spacecraft. The accelerometer had a resolution of better than 0.3 μGal.

The gradiometer is used in satellite missions (e.g., GOCE, Section 3.5.5) and in marine and airborne surveying. Several private companies have developed different proprietary techniques for measuring gravity gradients from an aircraft. The operating principle of one such gradiometer system, developed originally for the US military and modified for civilian use, is illustrated in Fig. 4.4. The gradiometer consists of sensitive accelerometers mounted in exactly aligned pairs on opposite rims of a disk that rotates slowly about a direction normal to the plane of the disk (e.g., the z-axis in Fig. 4.4). The difference between the measurements recorded by A1 and A2 gives the gravity gradient along the y-axis, at the same time as the second pair A3 and A4 measures the gradient along the x-axis. Each gradiometer repeats the other's measurement a quarter rotation later. The signal from each gradiometer is demodulated to obtain the gravitational gradient in the plane of rotation.

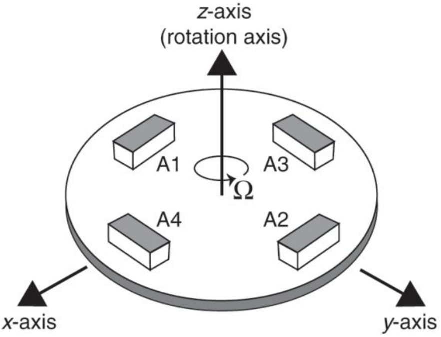


Fig. 4.4 Sketch showing the principle of operation of a gravitational gradiometer that employs two pairs of accelerometers (A1–A2 and A3–A4) mounted on a disk that rotates slowly about an axis normal to the disk.

In a full 3-D investigation of the gravitational gradients with this instrument, three disks carrying pairs of accelerometers rotate about orthogonal axes. This configuration is able to compensate for the accelerations experienced by the aircraft due to flight conditions. This kind of gravity gradient survey measures nine elements of the gravity field (Box 4.3) which form a symmetric matrix of values. The elements of the matrix are gravity gradients in the directions of specified geographic axes. Complex data processing provides a full solution of the gravity gradient tensor.

Box 4.3 The Gravity Gradient Matrix and Tensor

The orthogonal components (g_x, g_y, g_z) of the gravitational acceleration of a mass are related to the gravitational potential *U* of the mass by

$$g_x = -\frac{\partial U}{\partial x}; \qquad g_y = -\frac{\partial U}{\partial y}; \qquad g_z = -\frac{\partial U}{\partial z}. \tag{1}$$

The gradients of these accelerations in the direction of the *x*-axis are found by differentiating with respect to *x*:

$$\frac{\partial g_x}{\partial x} = -\frac{\partial^2 U}{\partial x^2}; \qquad \frac{\partial g_y}{\partial x} = -\frac{\partial^2 U}{\partial x \partial y} = \frac{\partial g_x}{\partial y}; \qquad \frac{\partial g_z}{\partial x} = -\frac{\partial^2 U}{\partial x \partial z} = \frac{\partial g_x}{\partial z}. \tag{2}$$

Note the equivalence of differentiating the *y*-component with respect to *x* and the *x*-component with respect to *y*. The same symmetry applies to differentiations with respect to *y* and *z*. For brevity we write the gradients as

$$\frac{\partial g_x}{\partial x} = G_{xx}; \qquad \frac{\partial g_x}{\partial y} = G_{xy}; \qquad \frac{\partial g_x}{\partial z} = G_{xz}. \tag{3}$$

Similar expressions are obtained by differentiating the gravity components with respect to *y* and *z*. The gradients form a *matrix* G with nine elements G_{ij}, whereby *i* and *j* successively take the values *x, y*, and *z* (Fig. B4.3.1). G_{ij} is a short form of representing the matrix elements, known as the gradient *tensor* with indices *i* and *j*. Tensor notation provides a concise form of algebra that can be used to express and solve relationships between physical vectors:

$$G = \begin{bmatrix} G_{xx} & G_{xy} & G_{xz} \\ G_{yx} & G_{yy} & G_{yz} \\ G_{zx} & G_{zy} & G_{zz} \end{bmatrix}. \tag{4}$$

Eq. (2) demonstrates equalities among the off-diagonal elements of the matrix that make it symmetric, such that

$G_{xy} = G_{yx}$, $G_{xy} = G_{yx}$, and $G_{xy} = G_{yx}$. In addition, the gravitational potential U must satisfy Laplace's equation, with the result that the sum of the diagonal elements (known as the trace of the matrix) is zero:

$$G_{xx} + G_{yy} + G_{zz} = 0. \qquad (5)$$

Consequently, only five elements of the matrix are independent.

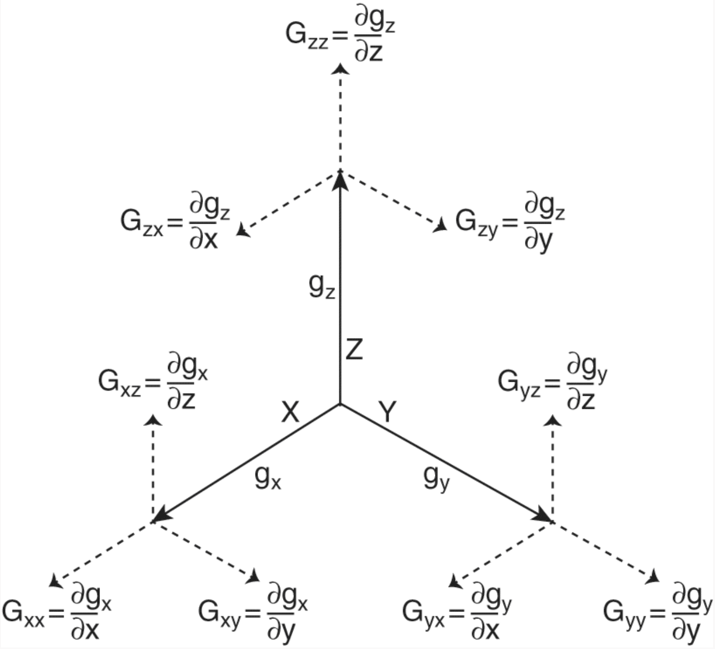

$G_{zz}=\frac{\partial g_z}{\partial z}$
$G_{zx}=\frac{\partial g_z}{\partial x}$
$G_{zy}=\frac{\partial g_z}{\partial y}$
g_z
Z
$G_{xz}=\frac{\partial g_x}{\partial z}$
$G_{yz}=\frac{\partial g_y}{\partial z}$
X
Y
g_x
g_y
$G_{xx}=\frac{\partial g_x}{\partial x}$
$G_{xy}=\frac{\partial g_x}{\partial y}$
$G_{yx}=\frac{\partial g_y}{\partial x}$
$G_{yy}=\frac{\partial g_y}{\partial y}$

Fig. B4.3.1 The orthogonal components of gravity (g_x, g_y, g_z) and their spatial gradients (G_{xx}, G_{yy}, G_{xy}, etc).

4.1.4 Gravity Survey Methods

The goal of gravity surveying is to locate and describe subsurface structures from the gravity effects caused by their anomalous densities. Most commonly, gravimeter measurements are made at a network of stations, spaced according to the purpose of the survey. In environmental studies a detailed high-resolution investigation of the gravity expression of a small area requires small distances of a few meters between measurement stations. In regional gravity surveys, as used for the definition of hidden structures of prospective commercial interest, the distance between stations may be several kilometers. If the area surveyed is not too large, a suitable site is selected as a base station (or reference site), and the gravity differences between the surveyed sites and this site are measured. In a gravity survey on a national scale, the gravity differences may be determined relative to a site where the absolute value of gravity is known.

4.1.4.1 Airborne Gravimetry

Marine and airborne surveys can measure variations in gravity and gravity gradients. Oceanographers have charted gravity anomalies over the oceans for decades, with important contributions to knowledge of ocean-bottom topography and the marine geoid. Airborne gravity measurement is subject to more severe inertial accelerations and was developed more recently. It is particularly useful for rapidly exploring uncharted areas on land, such as the many wilderness areas on each continent. As in magnetic surveying, parallel flight-lines are flown at a predetermined separation (Fig. 11.25), usually close to the topography to improve the resolution of anomalies. Comparisons with ground surveys show that similar resolution can be obtained in airborne surveys, but in a fraction of the measurement time.

There are two main methods for measuring gravity from an airplane or ship. One method uses a spring-type gravimeter of the Lacoste and Romberg type to measure relative gravity variations. The proof mass of the instrument is kept stationary on the spring suspension by an electromagnetic feedback system. The change in voltage needed to maintain balance is a measure of the relative change in gravity.

In the second type of airborne survey, different configurations of accelerometers are used to measure the gravitational field or its gradient. Accelerometers mounted on three slowly rotating disks give a full solution of the gravity gradient tensor. A simpler configuration with a single rotating disk allows partial solution of the tensor and delivers good agreement with land-based results. Proprietary algorithms filter and convert the data to the vertical component of gravity and its vertical gradient.

Two important sources of error must be taken into account in marine and airborne gravimetry. The first consists of inertial accelerations of the gravimeter that result from disturbances of the motion of the vehicle. In marine investigations these result from wave action, winds, and storms. Airborne gravimetry is affected by turbulence or air pockets, which arise from local low atmospheric pressure or downward winds and cause the airplane to drop abruptly. These motions are minimized by installing the gravimeter close to the roll center of the ship or airplane on a platform that is stabilized by gyroscopes and accelerometers. Stabilization of the measurement platform is an essential factor in airborne gravimetry, as is the on-board acquisition of accurate (e.g., dual-frequency) GPS data to monitor and compensate horizontal and vertical accelerations. Nevertheless, the motion effects can be very large and require removal by data processing.

The second source of error is the Eötvös effect due to the east–west component of the speed of the airplane or ship (Section 3.3.3). The position and speed of the moving vehicle are determined by continuously tracking its position with GPS observations. This gives accurate measures of the speed and heading of the aircraft and enables correction of the gravity measurements for the Eötvös effect.

4.1.5 Correction of Gravity Measurements

If the interior of the Earth were uniform, the value of gravity on the International Reference Ellipsoid would vary with latitude according to the normal gravity formula (Eq. 3.59). This provides us with a reference value for gravity measurements. In practice, it is not possible to measure gravity on the ellipsoid: the elevation of a measurement station may be hundreds of meters above or below the ellipsoid. Moreover, the gravity station may be surrounded by mountains and valleys that perturb the measurement. For example, let P and Q represent gravity stations at different elevations in hilly terrain (Fig. 4.5a). The theoretical value of gravity is computed at point R on the reference ellipsoid below P and Q. Thus, we must correct the measured gravity before it can be compared with the reference value.

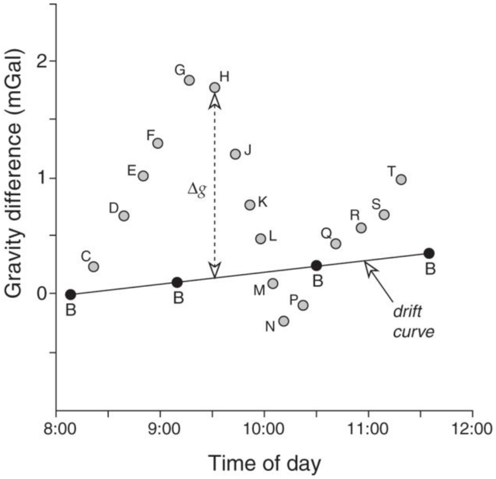


Fig. 4.6 Compensation of gravity readings for instrumental drift. The gravity stations B–T are occupied in sequence at known times. The repeated measurements at the base station B allow a drift correction to be made to the gravity readings at the other stations.

The hill to the right of stations P and Q has a center of mass that lies higher than the measurement elevation (Fig. 4.5a). The gravimeter measures gravity in the vertical direction, along the local plumb-line (see Fig. 3.6b). The mass of the hill-top above P attracts the gravimeter and causes an acceleration with a vertically upward component at P. The measured gravity is reduced by the presence of the hill-top; to compensate for this a *terrain* (or *topographic*) *correction* is calculated and *added* to the measured gravity. A similar effect is observed at Q, but the hill-top above Q is smaller and the corresponding terrain correction is smaller. These corrections effectively level the topography to the same elevation as the gravity station. The presence of a valley next to each measurement station also requires a terrain correction. In this case, imagine that we could fill the valley up to the level of each station with rock of the same density ρ as under P and Q. The downward attraction on the gravimeter would be increased, so the terrain correction for a valley must also be *added* to the measured gravity, just as for a hill. Removing the effects of the topography around a gravity station requires making positive terrain corrections (Δg_T) for both hills and valleys.

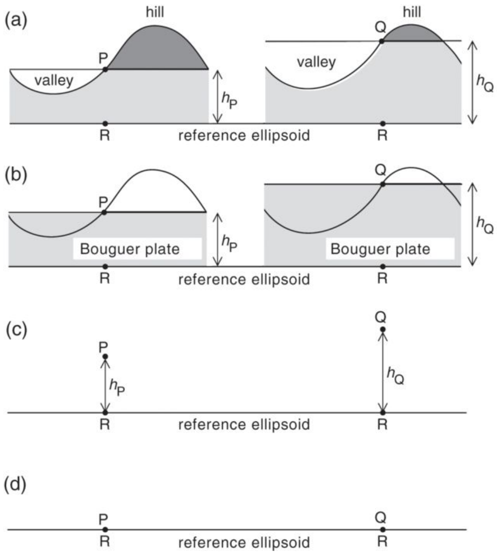

(a)
hill
hill
Q
P
valley
valley
h_P
h_Q
R
reference ellipsoid
R
(b)
Q
P
Bouguer plate
Bouguer plate
h_P
h_Q
R
reference ellipsoid
R
(c)
Q
P
h_P
h_Q
R
reference ellipsoid
R
(d)
P
Q
R
reference ellipsoid
R

Fig. 4.5 After (a) terrain corrections, (b) the Bouguer plate correction, and (c) the free-air correction, the gravity measurements at stations P and Q can be compared to the theoretical gravity at R

on the reference ellipsoid.

After effectively leveling the topography with the terrain correction, the rock between the gravity station and the reference ellipsoid is modeled as a uniform layer with density ρ (Fig. 4.5b). The gravitational acceleration of this rock-mass is included in the measured gravity and must be removed before we can compare with the theoretical gravity. The layer is taken to be a flat disk or plate of thickness h_P or h_Q under each station; it is called the *Bouguer plate*. Its gravitational acceleration can be computed for known thickness and density ρ, and gives a *Bouguer plate correction* (Δg_{BP}) that must be subtracted from the measured gravity, if the gravity station is above sea level. Note that if the gravity station is below sea level, we have to fill the space above it up to sea level with rock of density ρ; this requires increasing the measured gravity correspondingly. The Bouguer plate correction (Δg_{BP}) is negative if the station is above sea level but positive if it is below sea level. Its size depends on the density of the local rocks, but typically amounts to about 0.1 mGal m^{-1}.

Finally, we must compensate the measured gravity for the elevation h_P or h_Q of the gravity station above the ellipsoid (Fig. 4.5c). The main part of gravity is due to gravitational attraction, which decreases proportionately to the inverse square of distance from the center of the Earth. The gravity measured at P or Q is smaller than it would be if measured on the ellipsoid at R. A *free-air correction* (Δg_{FA}) for the elevation of the station must be added to the measured gravity. This correction ignores the effects of material between the measurement and reference levels, as this is taken care of in Δg_{BP}. Note that, if the gravity station were below sea level, the gravitational part of the measured gravity would be too large by comparison with the reference ellipsoid; we would need to subtract Δg_{FA} in this case. The free-air correction is positive if the station is above sea level but negative if it is below sea level (as might be the case in Death Valley or beside the Dead Sea). It amounts to about 0.3 mGal m^{-1}.

The free-air correction is always of opposite sense to the Bouguer plate correction. For convenience, the two are often combined in a single *elevation correction*, which amounts to about 0.2 mGal m^{-1}. This must be added for gravity stations above sea level and subtracted if gravity is measured below sea level. In addition, a *tidal correction* (Δg_{tide}) must be made (Section 3.3.4), and, if gravity is measured in a moving vehicle, the *Eötvös correction* (Section 3.3.3) is also necessary.

After correction the measured gravity can be compared with the theoretical gravity on the ellipsoid (Fig. 4.5d). Note that the above procedure reduces the measured gravity to the surface of the ellipsoid. In principle it is equally valid to correct the theoretical gravity from the ellipsoid upward to the level where the measurement was made. This method is preferred in more advanced types of analysis of gravity anomalies where the possibility of an anomalous mass between the ellipsoid and ground surface must be taken into account.

4.1.5.1 Time-Dependent Corrections

If a gravimeter is set up at a given place and monitored continuously, the readings are found to vary smoothly with time. The changes may amount to several hundredths of a mGal and thus may exceed the gravity anomaly that is to be measured. In part, the changes are a characteristic of the instrument; the remainder are caused by the tides. Transient changes are unrelated to the gravitational attraction of any local mass anomaly, therefore gravity measurements must be corrected for them.

The *instrumental drift* is partly due to thermally induced changes in the elastic properties of the gravimeter spring, which are minimized by housing the critical elements in an evacuated chamber. In addition, the elastic properties of the spring are not perfect, but creep slowly with time. The effect is small in modern gravimeters and can be compensated by making a *drift correction*. This is obtained by repeated occupation of some measurement stations at intervals during the day

(Fig. 4.6). Gravity readings at other stations are adjusted by comparison with the drift curve. In order to make this correction the time of each measurement must be noted.

During the course of a day, while measurements are being made at a particular location, the gravimeter is subject to tidal attractions, including vertical displacement due to the bodily Earth tides. The theory of the tides is known very well (see Section 3.3.4) and their time-dependent effects on gravity can be computed precisely for any place on Earth at any particular time. Consequently the *tidal correction* also requires that the time of each measurement be carefully noted.

4.1.5.2 Latitude Correction

The theoretical gravity at a given latitude is given by the normal gravity formula (Section 3.4.4, Eq. 3.59). If the measured gravity is an absolute measurement, the correction for latitude is made by subtracting the value predicted by this formula. Usually, however, a gravity survey is made with a gravimeter, and the quantity measured, g_m, is the gravity difference relative to a base station. The normal reference gravity g_n may then be replaced by a latitude correction, such as may be obtained by differentiating the approximate normal gravity formula, given by Eq. (3.59). Neglecting the β_2 term, which is three orders of magnitude smaller than the β_1 term, we get:

$$\frac{dg_n}{d\lambda} = g_e\beta_1 \sin 2\lambda. \tag{4.8}$$

After converting $d\lambda$ from radians to kilometers the latitude correction (Δg_{lat}) is found to be 0.8139 sin 2λ mGal per kilometer of north–south displacement. The correction is maximum at latitude 45°. Because gravity increases toward the poles, the correction for stations closer to the pole than the base station must be subtracted from the measured gravity, while for stations further from the pole the correction must be added.

4.1.5.3 Terrain Corrections

The terrain correction (Δg_T) for a hill adjacent to a gravity station is computed by dividing the hill into a number of vertical prisms (Fig. 4.7a). The contribution of each vertical element to the vertical acceleration at the point of observation P is calculated by assuming cylindrical symmetry about P. The height of the prism is h, its inner and outer radii are r_1 and r_2, respectively, the angle subtended at P is ϕ_0 (Fig. 4.7b), and the density of the hill is ρ. Let the sides of a small cylindrical element be dr, dz, and r dϕ; its mass is dm = ρ r dϕ dr dz and its contribution to the upward acceleration caused by the prism at P is

$$\Delta g = G\frac{\mathrm{d}m}{(r^2+z^2)}\cos\theta = G\frac{\rho r\,\mathrm{d}r\,\mathrm{d}z\,\mathrm{d}\phi}{(r^2+z^2)}\frac{z}{\sqrt{(r^2+z^2)}}. \quad (4.9)$$

Combining and rearranging terms and the order of integration gives the upward acceleration at P due to the cylindrical prism:

$$\Delta g_T = G\rho\int_{\phi=0}^{\phi_0}\mathrm{d}\phi\int_{r=r_1}^{r_2}\left(\int_{z=0}^{h}\frac{z\mathrm{d}z}{(r^2+z^2)^{3/2}}\right)r\mathrm{d}r. \quad (4.10)$$

The integration over ϕ gives ϕ_0; after further integration over z we get:

$$\Delta g_T = G\rho\phi_0\int_{r=r_1}^{r_2}\left(\frac{r}{\sqrt{(r^2+h^2)}}-1\right)\mathrm{d}r. \quad (4.11)$$

Integration over r gives the upward acceleration produced at P by the cylinder:

$$\Delta g_{\mathrm{T}} = G\rho\phi_0\left(\left(\sqrt{(r_1^2 + h^2)} - r_1\right) - \left(\sqrt{(r_2^2 + h^2)} - r_2\right)\right). \tag{4.12}$$

The direction of Δg_{T} in Fig. 4.7b is upward, opposite to gravity; the corresponding terrain correction must be *added* to the measured gravity.

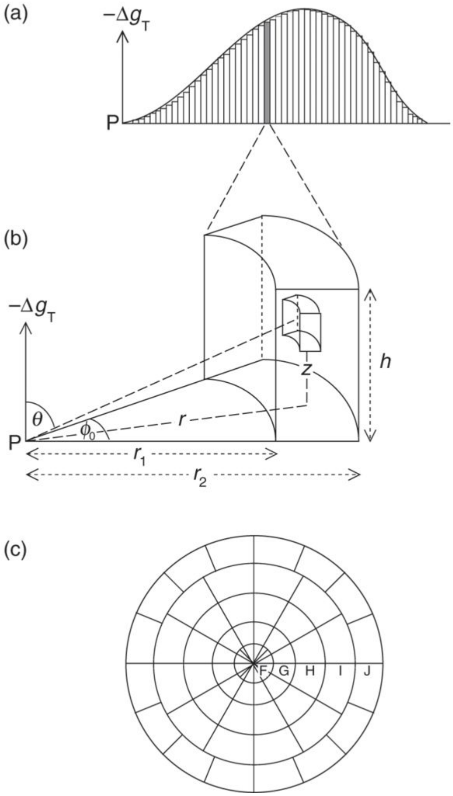


Fig. 4.7 Terrain corrections Δg_{T} are made by (a) dividing the topography into vertical elements, (b) computing the correction for each cylindrical element according to its height above or below the measurement station, and (c) adding up the contributions for all elements around the station with the aid of a transparent overlay on a topographic map.

Historically, terrain corrections were made using a terrain chart

(Fig. 4.7c) on which concentric circles and radial lines divide the area around the gravity station into sectors; these have radial symmetry like the cross-section of the element of a vertical cylinder in Fig. 4.7b. The inner and outer radius of each sector correspond to r_1 and r_2, and the angle subtended by the sector is ϕ. The terrain correction for each sector within each zone is pre-calculated using Eq. (4.12) and tabulated. The chart is drawn on a transparent sheet that is overlaid on a topographic map at the same scale and centered on the gravity station. The mean elevation within each sector is estimated as accurately as possible, and the elevation difference of the sector relative to the station is computed (i.e., h in Eq. 4.12). This is multiplied by the correction factor for the sector to give its contribution to the terrain correction. Finally, the terrain correction at the gravity station is obtained by summing up the contributions of all sectors. The procedure must be repeated for each gravity station. When the terrain chart is centered on a new station, the mean topographic relief within each sector changes and must be computed anew. The most important effects come from the topography nearest to the station. However, terrain corrections are generally necessary if a topographic difference within a sector is more than about 5% of its distance from the station.

The traditional method of making terrain corrections was both time-consuming and tedious. The computational principle is followed in modern analyses, which use databases of digitized topography instead of simple maps. As a result, the terrain corrections can now be made rapidly and easily by computers. For topography that is more than 22 km from the measurement station, the effect of the curvature of the Earth is also taken into account.

4.1.5.4 Bouguer Plate Correction

The Bouguer plate correction (Δg_{BP}) compensates for the effect of a layer of rock whose thickness corresponds to the elevation difference between the measurement and reference levels. This is modeled by a solid disk of density ρ and infinite radius centered at the gravity station P. The model is clearly an oversimplification. To obtain higher accuracy the curvature of the Earth must be taken into account. This is achieved by replacing the horizontal disk with a spherical cap of limited surface radius. Nevertheless, the simple Bouguer plate model suffices for many surveys.

The correction is computed by extension of the calculation for the terrain correction. An elemental cylindrical prism is defined as in Fig. 4.7b. Let the angle ϕ subtended by the prism increase to 2π and the inner radius decrease to zero; the first term in brackets in Eq. (4.12) reduces to *h*. The gravitational acceleration at the center of a solid disk of radius *r* is then

$$\Delta g_T = 2\pi G\rho\left(h - \left(\sqrt{(r^2 + h^2)} - r\right)\right). \tag{4.13}$$

Now let the radius *r* of the disk increase. The value of *h* gradually becomes insignificant compared to *r*; in the limit, when *r* is infinite, the second term in Eq. (4.13) tends to zero. Thus, the Bouguer plate correction (Δg_{BP}) is given by

$$\Delta g_{BP} = 2\pi G\rho h. \tag{4.14}$$

Inserting numerical values gives $4.19\rho \times 10^{-5}$ mGal m^{-1} for Δg_{BP}, where the density ρ is in kg m^{-3} (see Section 4.1.6). The correct choice of density is very important in computing Δg_{BP} and Δg_T. Some methods of determining the optimum choice are described in detail in Section 4.1.6.

An additional consideration is necessary in marine gravity surveys. Δg_{BP} requires uniform density below the surface of the reference ellipsoid. To compute Δg_{BP} over an oceanic region we must in effect replace the seawater with rock of density ρ. However, the measured gravity contains a component due to the attraction of the seawater (density 1030 kg m^{-3}) in the ocean basin. The Bouguer plate correction in marine gravity surveys is therefore made by replacing the density ρ in Eq. (4.14) by (ρ – 1030) kg m^{-3}. When a shipboard gravity survey is made over a large, deep lake, a similar allowance must be made for the depth of water in the lake using an assumed density of (ρ– 1000) kg m^{-3}.

4.1.5.5 Free-Air Correction

The free-air correction (Δg_{FA}) has a rather colorful, but slightly misleading title, giving the impression that the measurement station is floating in air above the ellipsoid. The density of air at standard temperature and pressure is around 1.3 kg m^{-3} and a mass of air between the observation and reference levels would cause a detectable gravity effect of about 50 µGal at an elevation of 1000 m. In fact, the free-air correction pays no attention to the density of the material between the measurement elevation and the ellipsoid. It is a straightforward correction for the decrease of gravitational acceleration with distance from the center of the Earth:

$$\frac{\partial g}{\partial r} = \frac{\partial}{\partial r}\left(-G\frac{E}{r^2}\right) = +2G\frac{E}{r^3} = -\frac{2}{r}g. \tag{4.15}$$

On substituting the Earth’s radius (6371 km) for *r* and the mean value of gravity (981,000 mGal) for *g*, the value of Δg_{FA} is found to be 0.3086 mGal m^{-1}.

4.1.5.6 Combined Elevation Correction

The free-air and Bouguer plate corrections are often combined into a single elevation correction, which is (0.3086 – (4.19ρ × 10^{-5})) mGal m^{-1}. Substituting a typical density for crustal rocks, usually taken to be 2670 kg m^{-3}, gives a combined elevation correction of

0.197 mGal m^{-1}. This must be added to the measured gravity if the gravity station is above the ellipsoid and subtracted if it is below.

The high sensitivity of modern gravimeters allows an achievable accuracy of 0.01–0.02 mGal in modern gravity surveys. To achieve this accuracy the corrections for the variations of gravity with latitude and elevation must be made very exactly. This requires that the precise coordinates of a gravity station must be determined by accurate geodetic surveying. The necessary precision of horizontal positioning is indicated by the latitude correction. This is maximum at 45° latitude, where, in order to achieve a survey accuracy of ± 0.01 mGal, the north–south positions of gravity stations must be known to about ± 10 m. The requisite precision in vertical positioning is indicated by the combined elevation correction of 0.2 mGal m^{-1}. To achieve a survey accuracy of ± 0.01 mGal the elevation of the gravimeter above the reference ellipsoid must be known to about ± 5 cm.

The elevation of a site above the ellipsoid is not the same as its altitude above mean sea level, which is equated with the geoid and not with the ellipsoid. Geoid undulations can amount to tens of meters relative to the ellipsoid (Section 3.4.5.1). However, they are long-wavelength features. Within a local survey the height difference between geoid and ellipsoid may not vary much, and the gravity differences from the selected base station are unlikely to be strongly affected. In surveys at a regional scale the discrepancies due to geoid undulations may be more serious. In the event that geoid undulations are large enough to affect a survey, the station altitudes must be corrected to true elevations above the ellipsoid.

4.1.6 Density Determination

The density of rocks in the vicinity of a gravity profile is important for the calculation of the Bouguer plate and terrain corrections. Density is defined as the mass per unit of volume of a material. It has different units and different numerical values in the c.g.s. and SI systems. For example, the density of water is 1 g cm^{-3} in the c.g.s. system, but 1000 kg m^{-3} in the SI system. In gravity prospecting, c.g.s. units are still in common use but are slowly being replaced by SI units. The formulae given for Δg_T and Δg_{BP} in Eqs. (4.12) and (4.14), respectively, require that density be given in kg m^{-3}.

A simple way of determining the appropriate density to use in a gravity study is to make a representative collection of rock samples with the aid of a geological map. The *specific gravity* of a sample may be found directly by weighing it first in air and then in water, and applying Archimedes' principle. This gives its density ρ_r relative to that of water:

$$\rho_r = \frac{W_a}{W_a - W_W}. \tag{4.16}$$

Typically, the densities found for different rock types by this method show a large amount of scatter about their means, and the ranges of values for different rock types overlap (Fig. 4.8). The densities of igneous and metamorphic rocks are generally higher than those of sedimentary rocks. This method is adequate for reconnaissance of an area. Unfortunately, it is often difficult to ensure that the surface collection of rocks is representative of the rock types in subsurface structures, so alternative methods of determining the appropriate density are usually employed. Density can be measured in vertical boreholes, drilled to explore the nature of a presumed structure. The density determined in the borehole is used to refine the interpretation of the structure.

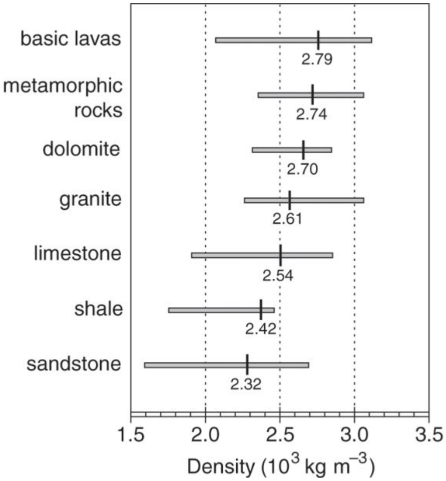


Fig. 4.8 Typical mean values and ranges of density for some common rock types.

Data source: M.B. Dobrin, *Introduction to Geophysical Prospecting* (3rd ed.), 630 pp., New York: McGraw-Hill, 1976.

4.1.6.1 Density from Seismic Velocities

Measurements on samples of water-saturated sediments and sedimentary rocks and on igneous and metamorphic rocks show that density and the seismic P-wave and S-wave velocities are related. The optimum fit to each dataset is a smooth curve (Fig. 4.9). Each curve is rather idealized, as the real data contain considerable scatter. For this reason the curves are best suited for computing the mean density of a large crustal body from its mean seismic velocity. Adjustments must be made for the higher temperatures and pressures at depth in the Earth, which affect both the density and the elastic parameters of rocks. However, the effects of high pressure and temperature can only be examined in laboratory experiments on small specimens. It is not known to what extent the results are representative of the *in-situ* velocity–density relationship in large crustal blocks.

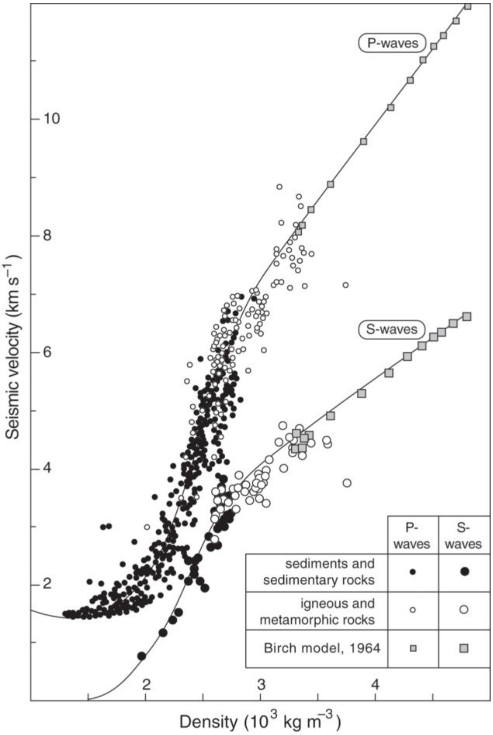


Fig. 4.9 The empirical relationships between density and the seismic P-wave and S-wave velocities in water-saturated sediments and sedimentary rocks and igneous and metamorphic rocks.

Source: W.J. Ludwig, J.E. Nafe and C.L. Drake, Seismic refraction, in: *The Sea*, Vol. 4, A.E. Maxwell, ed., pp. 53–84, New York: Wiley-Interscience, 1970.

The velocity–density curves are empirical relationships that do not have a theoretical basis. The P-wave data are used most commonly. In conjunction with seismic refraction studies, they have been used for modeling the density distributions in the Earth's crust and upper mantle responsible for large-scale, regional gravity anomalies (Section 4.2.4).

4.1.6.2 Gamma–Gamma Logging

The density of rock formations adjacent to a borehole can be determined from an instrument in the borehole. The principle makes use of the Compton scattering of γ-rays by loosely bound electrons in the rock adjacent to a borehole. An American physicist, Arthur H. Compton, discovered in 1923 that radiation scattered by loosely bound electrons experienced an increase in wavelength. This simple observation cannot be explained at all if the radiation is treated as a wave; the scattered radiation would have the same wavelength as the incident radiation. The Compton effect is easily explained by regarding the radiation as particles or photons, i.e., particles of quantized energy, rather than as waves. The energy of a photon is inversely proportional to its wavelength. The collision of a γ-ray photon with an electron is like that between billiard balls; part of the photon's energy is transferred to the electron. The scattered photon has lower energy and hence a longer wavelength than the incident photon. The Compton effect was an important verification of quantum theory.

The density logger, or *gamma–gamma* logger (Fig. 4.10), is a cylindrical device that contains a radioactive source of γ-rays, such as ^{137}Cs, which emits radiation through a narrow slit. The γ-ray photons collide with the loosely bound electrons of atoms near the hole, and are scattered. A scintillation counter to detect and measure the intensity of γ-rays is located in the tool about 45–60 cm above the emitter; the radiation reaching it also passes through a slit. Emitter and detector are shielded with lead, and the tool is pressed against the wall of the borehole by a strong spring, so that the only radiation registered is that resulting from the Compton scattering in the surrounding formation. The intensity of detected radiation is determined by the density of electrons, and so by the density of rock near to the logging tool. The γ-rays penetrate only about 15 cm into the rock.

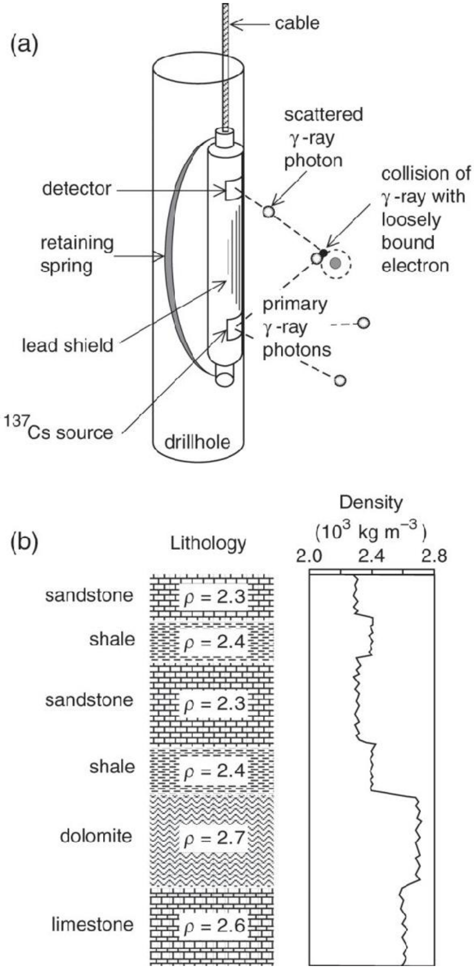

(a)
cable
scattered γ-ray photon
collision of γ-ray with loosely bound electron
detector
retaining spring
lead shield
primary γ-ray photons
^{137}Cs source
drillhole
(b)
Lithology
Density (10^3 kg m^{-3})
2.0
2.4
2.8
sandstone
ρ = 2.3
shale
ρ = 2.4
sandstone
ρ = 2.3
shale
ρ = 2.4
dolomite
ρ = 2.7
limestone
ρ = 2.6

Fig. 4.10 (a) The design of a gamma–gamma logging device for determining density in a borehole, and (b) a schematic gamma–gamma log calibrated in terms of the rock density.

Source: W.M. Telford, L.P. Geldart and R.E. Sheriff, *Applied Geophysics*, 770 pp., Cambridge: Cambridge University Press, 1990.

Calibrated gamma–gamma logs give the bulk density of the rock surrounding a borehole. This information is also needed for calculating porosity, which is defined as the fractional volume of the rock represented by pore spaces. Most sedimentary rocks are porous, the amount depending on the amount of compaction experienced. Igneous and metamorphic rocks generally have low porosity, unless they have been fractured. Usually the pores are filled with air, gas, or a fluid, such as water or oil. If the densities of the matrix rock and pore fluid are known, the bulk density obtained from gamma–gamma logging allows the porosity of the rock to be determined.

4.1.6.3 Borehole Gravimetry

Modern instrumentation allows gravity to be measured accurately in boreholes. One type of borehole gravimeter is a modification of the LaCoste–Romberg instrument, adapted for use in the narrow borehole and under conditions of elevated temperature and pressure. Alternative instruments have been designed on different principles; they have a comparable sensitivity of about 0.01 mGal. Their usage for down-hole density determination is based on application of the free-air and Bouguer plate corrections.

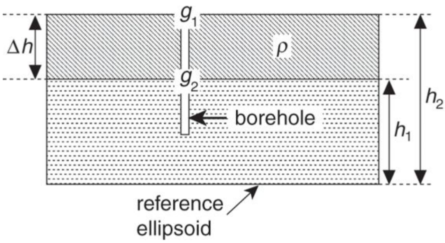


Fig. 4.11 Geometry for computation of the density of a rock layer from gravity measurements made in a vertical borehole.

Let g_1 and g_2 be the values of gravity measured in a vertical borehole at heights h_1 and h_2, respectively, above the reference ellipsoid (Fig. 4.11). The difference between g_1 and g_2 is due to the different heights and to the material between the two measurement levels in the borehole. The value g_2 will be larger than g_1 for two reasons. First, because the lower measurement level is closer to the Earth’s center, g_2 will be greater than g_1 by the amount of the combined elevation correction, namely $(0.3086 - (0.0419\rho \times 10^{-3}))\,\Delta h$ mGal, where $\Delta h = h_1 - h_2$. Second, at the lower level h_2 the gravimeter experiences an upward attraction due to the material between the two measurement levels. This reduces the measured gravity at h_2 and requires a compensating increase to g_2 of approximately $(0.0419\rho \times 10^{-3})\,\Delta h$ mGal. The difference Δg between the *corrected* values of g_1 and g_2 after reduction to the level h_2 is then

$$\begin{aligned}\Delta g &= (0.3086 - 0.0419\rho \times 10^{-3})\Delta h - 0.0419\rho \times 10^{-3}\Delta h \\ &= (0.3086 - 0.0838\rho \times 10^{-3})\Delta h.\end{aligned} \tag{4.17}$$

Rearranging this equation gives the density ρ of the material between the measurement levels in the borehole:

$$\rho = \left(3.683 - 11.93\frac{\Delta g}{\Delta h}\right) \times 10^3\ \mathrm{kg/m^3}. \tag{4.18}$$

If borehole gravity measurements are made with an accuracy of ±0.01 mGal at a separation of about 10 m, the density of the material near the borehole can be determined with an accuracy of about ±10 kg m^{-3}. More than 90% of the variation in gravity in the borehole is due to material within a radius of about $5\Delta h$ from the borehole (about 50 m for a distance $\Delta h \approx$ 10 m between measurement levels). This is much larger than the lateral range penetrated by gamma–gamma logging. As a result, effects related to the borehole itself are unimportant.

4.1.6.4 Nettleton's Method for Near-Surface Density

The near-surface density of the material under a hill can be determined by a method devised by L. Nettleton that compares the shape of a Bouguer gravity anomaly (Section 4.1.7) with the shape of the topography along a profile. The method makes use of the combined elevation correction ($\Delta g_{FA} + \Delta g_{BP}$) and the terrain correction (Δg_T), which are density-dependent. The terrain correction is less important than the Bouguer plate correction and can usually be neglected.

A profile of closely spaced gravity stations is measured across a small hill (Fig. 4.12). The combined elevation correction is applied to each measurement. Suppose that the true average density of the hill is 2600 kg m^{-3}. If the value assumed for ρ is too small (say, 2400 kg m^{-3}), Δg_{BP} at each station will be too small. The discrepancy is proportional to the elevation, so the Bouguer gravity anomaly is a *positive* image of the topography. If the value assumed for ρ is too large (say, 2800 kg m^{-3}), the opposite situation occurs. Too much is subtracted at each point, giving a computed anomaly that is a *negative* image of the topography. The optimum value for the density is found when the gravity anomaly has minimum correlation with the topography.

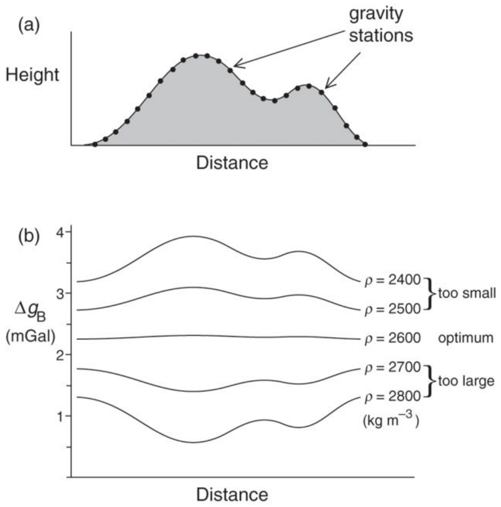


Fig. 4.12 Determination of the density of near-surface rocks by Nettleton’s method. (a) Gravity measurements are made on a profile across a small hill. (b) The data are corrected for elevation with various test values of the density. The optimum density gives minimum correlation between the gravity anomaly (Δg_B) and the topography.

4.1.7 Free-Air and Bouguer Gravity Anomalies

Suppose that we can measure gravity on the reference ellipsoid. If the distribution of density inside the Earth is homogeneous, the measured gravity should agree with the theoretical gravity given by the normal gravity formula (Eq. 3.59). The gravity corrections described in Section 4.1.5 compensate for the usual situation that the point of measurement is not on the ellipsoid. A discrepancy between the corrected, measured gravity and the theoretical gravity is called a *gravity anomaly*. It arises because the density of the Earth's interior is not homogeneous as assumed. The most common types of gravity anomaly are the *Bouguer anomaly* and the *free-air anomaly*.

The Bouguer gravity anomaly (Δg_B) is defined by applying all the corrections described individually in Section 4.1.5:

$$\Delta g_{\mathrm{B}} = g_{\mathrm{m}} + (\Delta g_{\mathrm{FA}} - \Delta g_{\mathrm{BP}} + \Delta g_{\mathrm{T}} + \Delta g_{\mathrm{tide}}) - g_{\mathrm{n}}. \tag{4.19}$$

In this formula g_m and g_n are the measured and normal gravity values; the corrections in parentheses are the free-air correction (Δg_{FA}), Bouguer plate correction (Δg_{BP}), terrain correction (Δg_T), and tidal correction (Δg_{tide}).

The free-air anomaly Δg_F is defined by applying only the free-air, terrain, and tidal corrections to the measured gravity:

$$\Delta g_{\mathrm{F}} = g_{\mathrm{m}} + (\Delta g_{\mathrm{FA}} + \Delta g_{\mathrm{T}} + \Delta g_{\mathrm{tide}}) - g_{\mathrm{n}}. \tag{4.20}$$

The Bouguer and free-air anomalies across the same structure can look quite different. Consider first the topographic block (representing a mountain range) shown in Fig. 4.13a. For this simple structure we neglect the terrain correction and assume tidal effects are zero. The difference between the Bouguer anomaly and the free-air anomaly arises from the Bouguer plate correction. In computing the Bouguer anomaly the simple elevation of the measurement station is taken into account together with the free-air correction. The measured gravity contains the attraction of the landmass above the ellipsoid, which is partially removed with the Bouguer plate correction. The underground structure does not vary laterally, so the corrected measurement agrees with the theoretical value and the Bouguer anomaly is everywhere zero across the mountain range. In computing the free-air anomaly only the free-air correction is applied; the part of the measured gravity due to the attraction of the landmass above the ellipsoid is not taken into account. Away from the mountain-block, the Bouguer and free-air anomalies are both equal to zero. Over the mountain the mass of the mountain-block increases the measured gravity compared to the reference value and results in a positive free-air anomaly across the mountain range.

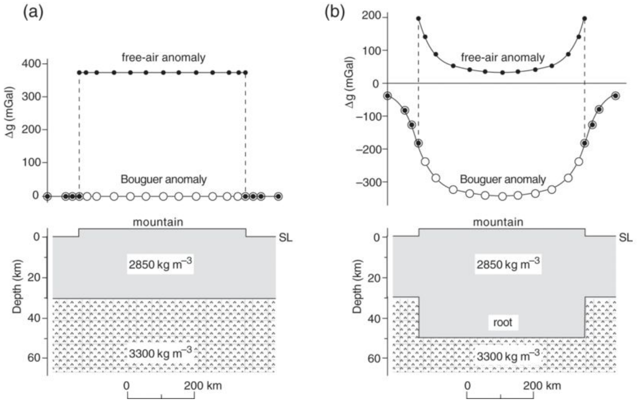


Fig. 4.13 Free-air and Bouguer anomalies across a mountain range. In (a) the mountain is modeled by a fully supported block, and in (b) the mass of the mountain above sea level (SL) is compensated by a less-dense crustal root, which projects down into the denser mantl.

Source: M.H.P. Bott, *The Interior of the Earth* (2nd ed.), London: Edward Arnold, 1982.

In fact, seismic data show that the Earth's crust is usually much thicker than normal under a mountain range. This means that a block of less-dense crustal rock projects down into the denser mantle (Fig. 4.13b). After making the free-air and Bouguer plate corrections there remains a Bouguer anomaly due to a block that represents the “root-zone” of the mountain range. As this is less dense than the adjacent and underlying mantle, it constitutes a mass deficit. The compensation of the mass of the mountain by a less-dense root-zone is called *isostasy*. The topic is treated in more detail in Sections 4.3, 5.4, and 5.5.

The attraction on a gravimeter at stations on a profile across the mountain range will be less than in Fig. 4.13a, so the corrected measurement will be less than the reference value. A strongly negative Bouguer anomaly is observed along the profile. At some distance from the mountain-block the Bouguer and free-air anomalies are equal but they are no longer zero, because the Bouguer anomaly now contains the effect of the root-zone. Over the mountain-block the free-air anomaly has a constant positive offset from the Bouguer anomaly, as in the previous example. Note that, although the free-air anomaly is positive, it falls to a very low value over the center of the block. At this point the attraction of the mountain is partly canceled by the missing attraction of the less-dense root-zone.

4.2 Interpretation of Gravity Anomalies

4.2.1 Regional and Residual Anomalies

A gravity anomaly results from the inhomogeneous distribution of density in the Earth. Suppose that the density of rocks in a subsurface body is ρ and the density of the rocks surrounding the body is ρ_0. The difference $\Delta\rho = \rho - \rho_0$ is called the *density contrast* of the body with respect to the surrounding rocks. If the body has a higher density than the host rock, it has a positive density contrast; a body with lower density than the host rock has a negative density contrast. Over a high-density body the measured gravity is augmented; after reduction to the reference ellipsoid and subtraction of the normal gravity, a positive gravity anomaly is obtained. Likewise a negative anomaly results over a region of low density. The presence of a gravity anomaly indicates a body or structure with anomalous density; the sign of the anomaly is the same as that of the density contrast and shows whether the density of the body is higher or lower than normal.

The appearance of a gravity anomaly is affected by the dimensions, density contrast, and depth of the anomalous body. The horizontal extent of an anomaly is often called its apparent "wavelength." The wavelength of an anomaly is a measure of the depth of the anomalous mass. Large, deep bodies give rise to broad (long-wavelength) anomalies, while small, shallow bodies cause narrow (short-wavelength) anomalies.

Usually a map of Bouguer gravity anomalies contains superposed anomalies from several sources. The long-wavelength anomalies due to deep density contrasts are called *regional* anomalies. They are important for understanding the large-scale structure of the Earth's crust under major geographic features, such as mountain ranges, oceanic ridges and subduction zones. Short-wavelength *residual* anomalies are due to shallow anomalous masses that may be of interest for commercial exploitation. Geological knowledge is essential for interpreting the residual anomalies. In eroded continental shield areas, like Canada or Scandinavia, anomalies with very short wavelengths may be due to near-surface mineralized bodies. In sedimentary basins, short- or intermediate-wavelength anomalies may arise from structures related to reservoirs for petroleum or natural gas.

4.2.2 Separation of Regional and Residual Anomalies

Separating the regional (long-wavelength) anomaly from the local residual (shorter-wavelength) anomaly is an important step in the interpretation of a gravity map. The analysis may be based on selected profiles across some structure, or it may involve the two-dimensional distribution of anomalies in a gravity map. Numerous techniques have been applied to the decomposition of a gravity anomaly into its constituent parts. They range in sophistication from simple visual inspection of the anomaly pattern to advanced mathematical analysis. A few examples of these methods are described in the following subsections.

4.2.2.1 Visual Analysis

The simplest way of representing the regional anomaly on a gravity profile is by visually fitting the large-scale trend with a smooth curve (Fig. 4.14). The value of the regional gravity given by this trend is subtracted point by point from the Bouguer gravity anomaly. This method allows the interpreter to visually fit curves that leave residual anomalies with a sign appropriate to his interpretation of the density distribution.

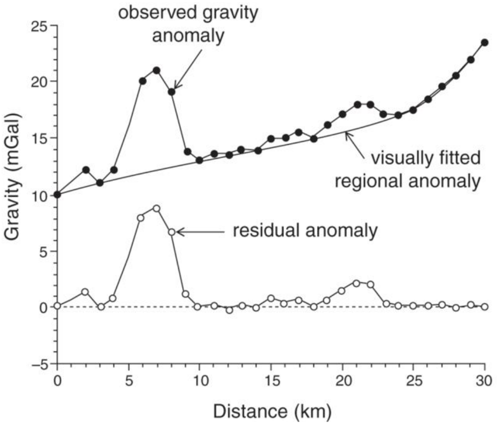


Fig. 4.14 Representation of the regional anomaly on a gravity profile by visually fitting the large-scale trend with a smooth curve.

This approach may be adapted to the analysis of a gravity map by visually smoothing the contour lines. In Fig. 4.15a the contour lines of equal Bouguer gravity curve sharply around a local abnormality. The more gently curved contours have been continued smoothly as dashed lines. They indicate how the interpreter thinks the regional gravity field (Fig. 4.15b) would continue in the absence of the local abnormality. The values of the regional and original Bouguer gravity anomalies are interpolated from the corresponding maps at points spaced on a regular grid. The regional value is subtracted from the Bouguer anomaly at each point and the computed residuals are contoured to give a map of the local gravity anomaly (Fig. 4.15c). The experience and skill of the interpreter are important factors in the success of visual methods.

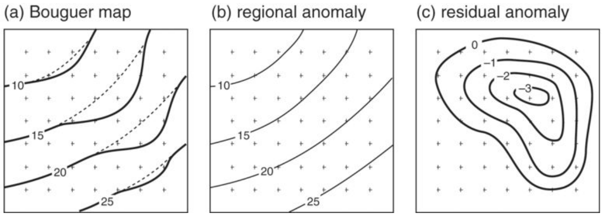


Fig. 4.15 Removal of regional trend from a gravity map by contour smoothing: (a) hand-drawn smoothing of contour lines on original Bouguer gravity map; (b) map of regional gravity variation; and (c) residual gravity anomaly after subtracting the regional variation from the Bouguer gravity map. Values are in mGal.

Source: E.S. Robinson and C. Çoruh, *Basic Exploration Geophysics*, 562 pp., New York: John Wiley, 1988

4.2.2.2 Polynomial Representation

In an alternative method, the regional trend is represented by a straight line, or, more generally, by a smooth polynomial curve. If x denotes the horizontal position on a gravity profile, the regional gravity Δg_R may be written as

$$\Delta g_R = \Delta g_0 + \Delta g_1 x + \Delta g_2 x^2 + \Delta g_3 x^3 + \cdots + \Delta g_n x^n.$$

(4.21)

The polynomial is fitted by the method of least squares to the observed gravity profile. This gives optimum values for the coefficients Δg_n. The method also has drawbacks. The higher the degree of the polynomial, the better it fits the observations (Fig. 4.16). The ludicrous extreme is when the degree of the polynomial is one less than the number of observations; the curve then passes perfectly through all the data points, but the regional gravity anomaly has no meaning geologically. The interpreter's judgment is important in selecting the degree of the polynomial, which is usually chosen to be the lowest possible degree that represents most of the regional trend. Moreover, a curve fitted by least squares must pass through the mean of the gravity values, so that the residual anomalies are divided equally between positive and negative values. Each residual anomaly is flanked by anomalies of opposite sign (Fig. 4.16), which are due to the same anomalous mass that caused the central anomaly and so have no significance of their own.

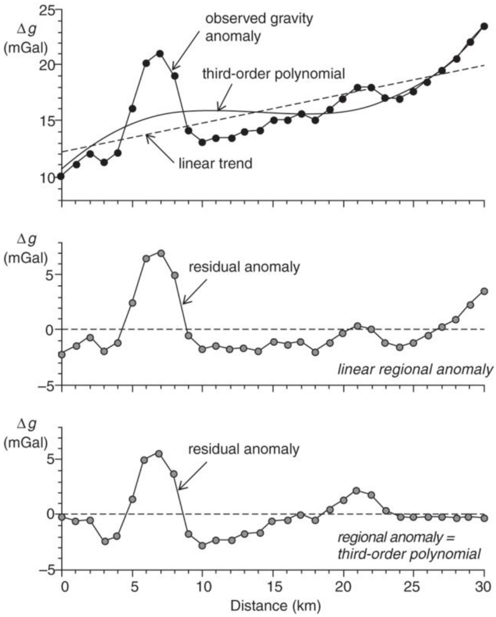


Fig. 4.16 Representation of the regional trend by a smooth

polynomial curve fitted to the observed gravity profile by the method of least squares.

Polynomial fitting can also be applied to gravity maps. It is assumed that the regional anomaly can be represented by a smooth surface, $\Delta g(x, y)$, which is a low-degree polynomial of the horizontal position coordinates x and y. In the simplest case the regional anomaly is expressed as a first-degree polynomial, or plane. To express changes in the gradient of gravity, a higher-degree polynomial is needed. For example, the regional gravity given by a second-degree polynomial is written as

$$\Delta g(x,y) = \Delta g_0 + \Delta g_{x1}x + \Delta g_{y1}y + \Delta g_{x2}x^2 + \Delta g_{y2}y^2 + \Delta g_{xy}xy. \tag{4.22}$$

As in the analysis of a profile, the optimum values of the coefficients Δg_{x1}, Δg_{y1}, etc., are determined by least-squares fitting. The residual anomaly is again computed point by point by subtracting the regional from the original data.

4.2.2.3 Representation of Gravity Anomalies by Fourier Series

The gravity anomaly along a profile can be analyzed with techniques developed for investigating time series. Instead of varying with time, as the seismic signal does in a seismometer, the gravity anomaly $\Delta g(x)$ varies with position x along the profile. For a spatial distribution the wave number, $k = 2\pi/\lambda$, is the counterpart of the frequency of a time series. If it can be assumed that its variation is periodic, the function $\Delta g(x)$ can be expressed as the sum of a series of discrete harmonics. Each harmonic is a sine or cosine function whose argument is a multiple of the fundamental wave number. The expression for $\Delta g(x)$ is called a *Fourier series* (Box 4.4).

Box 4.4 Fourier Analysis

If integrated over a full cycle, the resulting value of a sine or cosine function is zero. The integrated value of the product of a sine function and a cosine function is also zero. Mathematically, this defines the sines and cosines as *orthogonal* functions. The squared values of sines and cosines do not integrate to zero over a full cycle; this property can be used to *normalize* functions that can be expressed in terms of sines and cosines. These observations may be summarized as follows for the functions sin($n\theta$) and cos($n\theta$):

$$\begin{aligned}
&\int_0^{2\pi} \sin(n\theta)\mathrm{d}\theta = \int_0^{2\pi} \cos(n\theta)\mathrm{d}\theta = 0\\
&\int_0^{2\pi} \sin^2(n\theta)\mathrm{d}\theta = \frac{1}{2}\int_0^{2\pi}\Big(1 - \cos(2n\theta)\Big)\mathrm{d}\theta = \pi\\
&\int_0^{2\pi} \cos^2(n\theta)\mathrm{d}\theta = \frac{1}{2}\int_0^{2\pi}\Big(1 + \cos(2n\theta)\Big)\mathrm{d}\theta = \pi.
\end{aligned}$$

(1)

If follows by applying these results and invoking the formulae for the sums and differences of sines and cosines that

$$\begin{aligned}
&\int_0^{2\pi} \sin(n\theta)\cos(m\theta)\mathrm{d}\theta = \frac{1}{2}\int_0^{2\pi}\Big\{\sin\Big(\frac{n+m}{2}\theta\Big)\\
&\quad + \sin\Big(\frac{n-m}{2}\theta\Big)\Big\}\mathrm{d}\theta = 0\\
&\int_0^{2\pi} \cos(n\theta)\cos(m\theta)\mathrm{d}\theta = \int_0^{2\pi} \sin(n\theta)\sin(m\theta)\mathrm{d}\theta\\
&\quad = \frac{1}{2}\int_0^{2\pi}\Big\{\cos\Big(\frac{n-m}{2}\theta\Big) \pm \cos\Big(\frac{n+m}{2}\theta\Big)\Big\}\mathrm{d}\theta\\
&\quad = \left\{\begin{matrix} 0, & \text{if} \quad m \neq n\\ \pi, & \text{if} \quad m = n \end{matrix}\right\}.
\end{aligned}$$

(2)

These relationships form the basis of Fourier analysis.

Geophysical signals that vary periodically in time (e.g., seismic waves) or space (e.g., gravity, magnetic, or thermal anomalies) can be expressed as the superposition of harmonics of a fundamental frequency or wave number. For example, consider a gravity anomaly $\Delta g(x)$ along a profile of length L in the x-direction. The fundamental "wavelength" λ of the anomaly is equal to $2L$, i.e., twice the profile length. A corresponding fundamental wave number is defined for the profile as $k = (2\pi/\lambda)$, so that the argument θ in the sine and cosine functions in Eqs. (1) and (2) is replaced by (kx). The observed anomaly is represented by adding together components corresponding to harmonics of the fundamental wavelength:

$$\begin{aligned}\Delta g(x) &= a_0 + a_1 \cos(kx) + b_1 \sin(kx) + a_2 \cos(2kx) \\ &\quad + b_2 \sin(2kx) \ldots, \\ \Delta g(x) &= \sum_{n=0}^{N} \Big(a_n \cos(nkx) + b_n \sin(nkx) \Big),\end{aligned} \tag{3}$$

This expression for $\Delta g(x)$ is called a *Fourier series*. In Eq. (3) the summation is truncated after N sine and cosine terms. The value of N is chosen to be as large as necessary to describe the gravity anomaly adequately. The importance of any individual term of order n is given by the values of the corresponding coefficients a_n and b_n, which act as weighting functions.

The coefficients a_n and b_n can be calculated by using the orthogonal properties of sine and cosine functions summarized in Eqs. (1) and (2). If both sides of Eq. (3) for the gravity anomaly $\Delta g(x)$ are multiplied by $\cos(mkx)$, we get

$$\begin{aligned}\Delta g(x)\cos(mkx) = \sum_{n=0}^{N} \Big(& a_n \cos(nkx)\cos(mkx) \\ & + b_n \sin(nkx)\cos(mkx) \Big).\end{aligned}$$

(4)

Each product on the right-hand side of this equation can be written as the sum or difference of two sines or cosines. Thus,

$$\Delta g(x)\cos(mkx) = \frac{1}{2}\sum_{n=0}^{N}\Big(a_n[\cos((n+m)kx) + \cos((n-m)kx)] + b_n[\sin((n+m)kx) + \sin((n-m)kx)]\Big).$$

(5)

Integration of $\Delta g(x)$ cos($2mk$ x) over a full wavelength of x causes all terms on the right side of Eq. (5) to vanish unless $n = m$, when the term $\cos((n-m)kx) = \cos(0) = 1$. This allows us to determine the coefficients a_n in Eq. (2):

$$\int_0^\lambda \Delta g(x)\cos(nkx)dx = \frac{1}{2}a_n\int_0^\lambda dx = \frac{\lambda}{2}a_n,$$

(6)

$$a_n = \frac{2}{\lambda}\int_0^\lambda \Delta g(x)\cos(nkx)dx = \frac{1}{\pi}\int_0^{2\pi}\Delta g(x)\cos(n\theta)d\theta.$$

(7)

The coefficients b_n in Eq. (3) are obtained similarly by multiplying $\Delta g(x)$ by sin(mkx) and integrating over a full cycle of the signal. This gives

$$b_n = \frac{2}{\lambda}\int_0^\lambda \Delta g(x)\sin(nkx)dx = \frac{1}{\pi}\int_0^{2\pi}\Delta g(x)\sin(n\theta)d\theta.$$

(8)

The breakdown of a complex anomaly (or time series) in terms of simpler periodic variations of different wavelengths is called *Fourier analysis* and is a powerful method for resolving the most important components of the original signal.

The two-dimensional variation of a mapped gravity anomaly can be expressed in a similar way with the aid of *double Fourier series* (Box 4.5). As in the simpler one-dimensional case of a gravity anomaly on a profile, the expression of two-dimensional gravity anomalies by double Fourier series is analogous to summing weighted sinusoidal functions. These can be visualized as corrugations of the *x*–*y* plane (Fig. 4.17), with each corrugation weighted according to its importance to $\Delta g(x, y)$.

Box 4.5 Double Fourier Series

The two-dimensional variation of a mapped gravity anomaly can be analyzed with the aid of *double Fourier series*. In this case the gravity anomaly is a function of both the *x*- and *y*-coordinates and can be written

$$\Delta g(x,y) = \sum_{n=0}^{N}\sum_{m=0}^{M}(a_{nm}C_nC_m^* + b_{nm}C_nS_m^* + c_{nm}S_nC_m^* + d_{nm}S_nS_m^*), \quad (1)$$

where

$$C_n = \cos\left(2n\pi\frac{x}{\lambda_x}\right) \quad S_n = \sin\left(2n\pi\frac{x}{\lambda_x}\right)$$
$$C_m^* = \cos\left(2m\pi\frac{y}{\lambda_y}\right) \quad S_m^* = \sin\left(2m\pi\frac{y}{\lambda_y}\right). \quad (2)$$

In these expressions the fundamental wavelengths λ_x and λ_y express the extent of the anomaly in the *x*- and *y*-directions, respectively. The derivation of the coefficients a_{nm}, b_{nm}, c_{nm}, and d_{nm} is similar in principle to the one-dimensional case, relying on the orthogonality of the individual sine and cosine terms, and the property that the products of two sine terms, two cosine terms, or a sine term with a cosine term, can be expressed as the sum or difference of other sine or cosine functions. As might be expected, the analysis of double Fourier series is somewhat more complicated than in the one-dimensional case, but it delivers results that characterize the two-dimensional variation of the regional gravity anomaly.

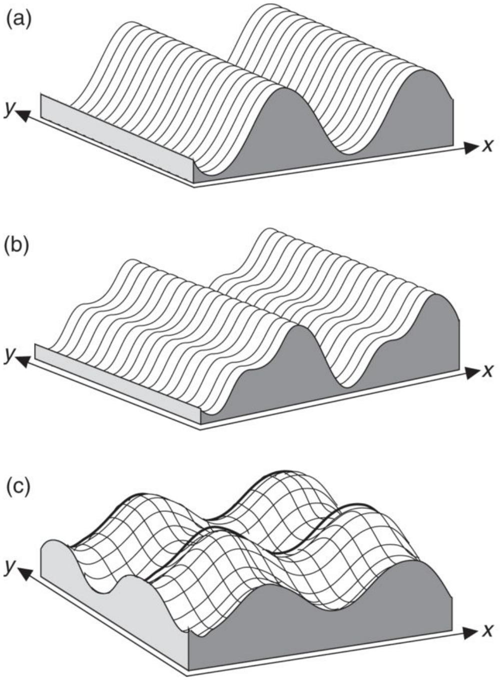


Fig. 4.17 Expression of the two-dimensional variation of a gravity anomaly using double Fourier series: (a) a single harmonic in the x-direction; (b) two harmonics in the x-direction; and (c) superposed single harmonics in the x- and y-directions, respectively.

Source: J.C. Davis, *Statistics and Data Analysis in Geology*, 550 pp., New York: John Wiley, 1973.

4.2.2.4 Anomaly Enhancement and Filtering

The above discussion shows how a function that is periodic can be expressed as a Fourier sum of harmonics of a fundamental wavelength. By breaking down the observed signal into discrete components, it is possible to remove some of these and reconstruct a filtered version of the original anomaly. However, the requirements of periodic behavior and discreteness of harmonic content are often not met. For example, the variation of gravity from one point to another is usually not periodic. Moreover, if the harmonic content of a function is made up of distinct multiples of a fundamental frequency or wave number, the wavelength spectrum consists of a number of distinct values. Yet many functions of geophysical interest are best represented by a continuous spectrum of wavelengths.

Box 4.6 Complex Numbers

In determining the roots of cubic and higher-order polynomials it is sometimes necessary to use the square roots of negative numbers. A negative number can be written as the positive number multiplied by (–1), so the square root of a negative number contains the imaginary unit i, defined as the square root of (–1). A complex number z consists of a real part x and an imaginary part y, and is written

$$z = x + \mathrm{i}y. \tag{1}$$

The number $z^* = x - \mathrm{i}y$ is called the *complex conjugate* of z. Note that the imaginary term has the same value as in Eq. (1) but has the opposite sign. The product $zz^* = x^2 + y^2$ gives the squared amplitude of the complex number.

The Swiss mathematician Leonhard Euler showed in 1748 how the trigonometric functions, $\cos\theta$ and $\sin\theta$, are related to a complex exponential function:

$$\cos\theta + \mathrm{i}\sin\theta = e^{\mathrm{i}\theta}, \tag{2}$$

where e is the base of the natural logarithms and i is the imaginary unit. This formula is fundamental to the study of functions of a complex variable, a branch of mathematics known as *complex analysis* that can be used to solve a wide range of practical and theoretical problems. In this expression, $\cos\theta$ is called the real part of $e^{\mathrm{i}\theta}$ and $\sin\theta$ is called the imaginary part. It follows that

$$\begin{aligned} \mathrm{i} &= e^{\mathrm{i}\pi/2} \\ \sqrt{\mathrm{i}} &= e^{\mathrm{i}\pi/4} = \frac{1}{\sqrt{2}}(1+\mathrm{i}). \end{aligned} \tag{3}$$

Complex numbers can be depicted geometrically on an *Argand diagram* (also called the *complex plane*), invented in 1806 by another Swiss mathematician, Jean-Robert Argand, as an aid to their visualization. In this diagram (Fig. B4.6.1) the real part of the number, x in our case, is plotted on the horizontal axis and the imaginary part, y, on the vertical axis. If the distance of point P at (x, y) from the origin of the plot at O is r, and the angle between OP and the x-axis is θ, then the real part of the complex number is $x = r\cos\theta$ and the imaginary part is $y = r\sin\theta$, so that, using Euler's formula,

$$z = x + \mathrm{i}y = r(\cos\theta + \mathrm{i}\sin\theta) = re^{\mathrm{i}\theta}. \tag{4}$$

Complex numbers are combined in the same way as real numbers. Two complex numbers, z_1 and z_2, combine to form a new complex number, z_3. A special case is when the complex number is combined with its *complex conjugate*; this results in a real number. Some examples of possible combinations of complex numbers are as follows:

$$\begin{aligned} z_3 &= z_1 + z_2 = (x_1 + x_2) + \mathrm{i}(y_1 + y_2) \\ z_3 &= r_1 e^{\mathrm{i}\theta_1} + r_2 e^{\mathrm{i}\theta_2}, \end{aligned} \tag{5}$$

$$\begin{aligned} z_3 &= z_1 z_2 = (x_1 + \mathrm{i}y_1)(x_2 + \mathrm{i}y_2) \\ &= (x_1 x_2 - y_1 y_2) + \mathrm{i}(x_1 y_2 + y_1 x_2) \\ z_3 &= (r_1 e^{\mathrm{i}\theta_1})(r_2 e^{\mathrm{i}\theta_2}) = r_1 r_2 e^{i(\theta_1 + \theta_2)}. \end{aligned} \tag{6}$$

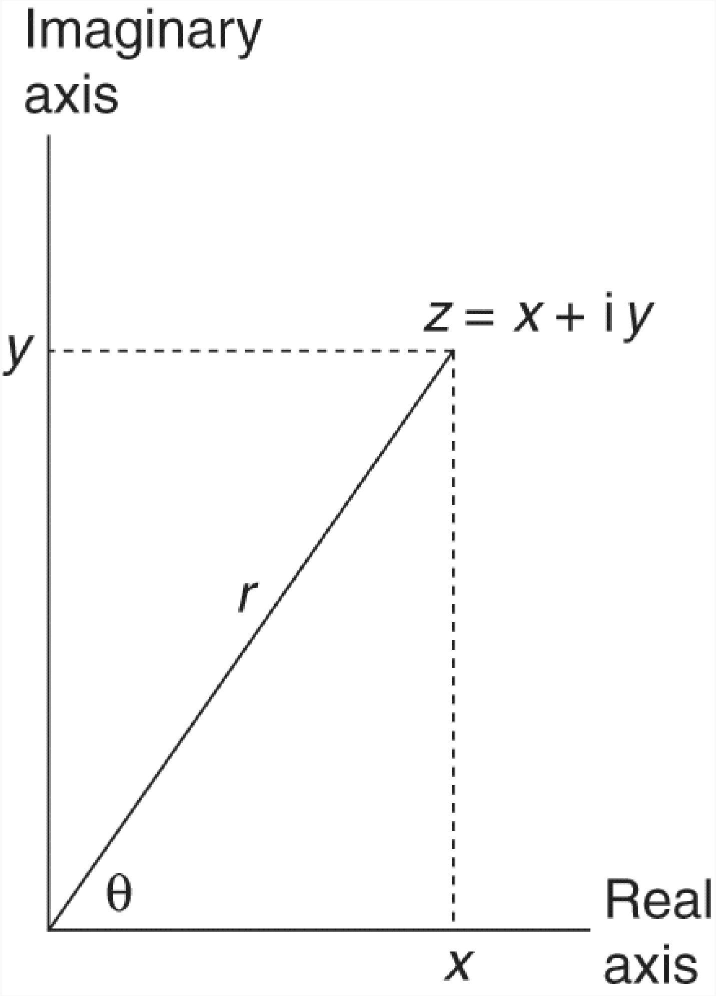


Fig. B4.6.1 The complex plane (or Argand diagram).

To handle this kind of problem the spatial variation of gravity is represented by a *Fourier integral*, which consists of a continuous set of frequencies or wave numbers instead of a discrete set. The Fourier integral can be used to represent non-periodic functions. It uses complex numbers (Box 4.6), which are numbers that involve *i*, the square root of –1. If the gravity anomaly is analyzed in two-dimensions (instead of on a single profile), a two-dimensional integral is needed, analogous to the double Fourier series representation described in Section 4.2.2.3 and Box 4.5. The observed gravity can then be manipulated using the techniques of *Fourier transforms*. These techniques involve intensive computations and are ideally suited to digital data processing with computers.

Box 4.7 Fourier Transforms

Functions that cannot be expressed as the Fourier series of individual terms, as in Boxes 4.5 and 4.6, may be replaced by a *Fourier integral.* This consists of a continuous set of frequencies or wave numbers instead of a discrete set. The Fourier integral can be used to represent non-periodic functions. The gravity anomaly $\Delta g(x)$ is now written as an integral instead of as a sum of discrete terms:

$$\Delta g(x) = \int_{-\infty}^{\infty} G(u) e^{iux} \mathrm{d}u, \tag{1}$$

where, by using the properties of complex numbers, it can be shown that

$$G(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Delta g(x) e^{-iux} \mathrm{d}x. \tag{2}$$

The complex function $G(u)$ is the *Fourier transform* of the real-valued function $\Delta g(x)$. An adequate treatment of Fourier transforms is beyond the scope of this book. However, the use of this powerful mathematical technique can be illustrated without delving deeply into the theory.

A map of gravity anomalies can be represented by a function $\Delta g(x, y)$ of the Cartesian map coordinates. The Fourier transform of $\Delta g(x, y)$ is a two-dimensional complex function that involves wave numbers k_x and k_y defined by wavelengths of the gravity field with respect to the x- and y-axes ($k_x = 2\pi/\lambda_x$, $k_y = 2\pi/\lambda_y$). It is

$$G(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Delta g(x,y)\{\cos(k_x x + k_y y) - \mathrm{i}\sin(k_x x + k_y y)\}\,\mathrm{d}x\,\mathrm{d}y. \quad (3)$$

This equation assumes that the observations $\Delta g(x, y)$ can be represented by a continuous function defined over an infinite x–y plane, whereas in fact the data are of finite extent and are known at discrete points of a measurement grid. In practice, these inconsistencies are usually not important. Efficient computer algorithms permit the rapid computation of the Fourier transform $G(x, y)$ of the gravity anomaly $\Delta g(x, y)$. The transformed signal can be readily manipulated in the Fourier domain by convolution with a filter function and the result deconvoluted back into the spatial domain. This allows the operator to examine the effects of low- and high-pass filtering on the observed anomaly, which can greatly help in its interpretation.

The two-dimensional Fourier transform of a gravity map makes it possible to digitally filter the gravity anomalies. A filter is a spatial function of the coordinates x and y. When a filter is applied to the function $\Delta g(x, y)$ representing the gravity data, the output is a map of the filtered gravity data. The computation in the spatial domain defined by the x- and y-coordinates can be time-consuming. It is often faster to compute the Fourier transforms of the gravity and filter functions, multiply these together in the Fourier domain, then perform an inverse Fourier transform on the product to convert it back to the spatial domain.

The nature of the filter applied in the Fourier domain can be chosen to eliminate certain wavelengths. For example, it can be designed to cut out all wavelengths shorter than a selected wavelength and to pass longer wavelengths. This is called a *low-pass filter*; it passes or preserves long wavelengths that have low wave numbers, removing the part of the anomaly having short wavelengths. The irregularities in a Bouguer gravity anomaly map (Fig. 4.18a) are removed by low-pass filtering, leaving a filtered map (Fig. 4.18b) that is much smoother than the original. Alternatively, the filter in the Fourier domain can be designed to eliminate wavelengths longer than a selected wavelength and to pass (or preserve) shorter wavelengths. The application of such a *high-pass filter* enhances the short-wavelength (high wave number) component of the gravity map (Fig. 4.18c).

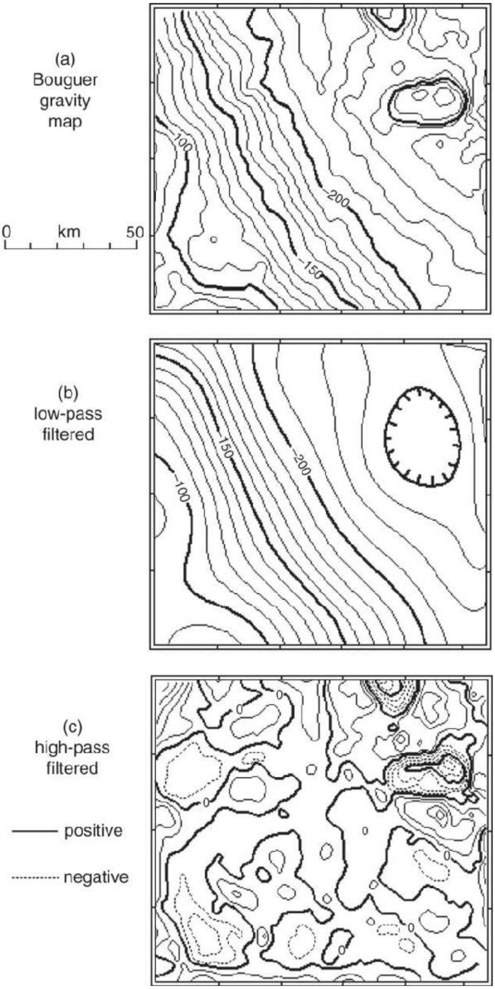

(a) Bouguer gravity map
0 km 50
100
200
150
(b) low-pass filtered
150
200
100
(c) high-pass filtered
positive
negative
0

Fig. 4.18 The use of wavelength filtering to emphasize selected anomalies in the Sierra Nevada, California: (a) unfiltered Bouguer gravity map; (b) low-pass filtered gravity map with long-wavelength regional anomalies; and (c) high-pass filtered gravity map enhancing short-wavelength local anomalies. Contour interval: (a) and (b) 10 mGal, (c) 5 mGal.

Source: M.B. Dobrin and C.H. Savit, *Introduction to Geophysical Prospecting* (4th ed.), New York: McGraw-Hill, 1988.

Wavelength filtering can be used to emphasize selected anomalies. For example, in studying large-scale crustal structure the gravity anomalies due to local small bodies are of less interest than the regional anomalies, which can be enhanced by applying a low-pass filter. Conversely, in the investigation of anomalies due to shallow crustal sources the regional effect can be suppressed by high-pass filtering.

4.2.2.5 Upward and Downward Continuation

In some cases it is desirable to know the gravity anomalies on a different surface from that on which the measurements were made. Provided no additional masses lie between the surfaces, Laplace's equation (Box 3.4) is valid in the intervening space. Using first and second derivatives of the measured gravity map, the anomalies can be computed on the new surface. If this is at a higher altitude than the measurement surface, the computational process is called *upward continuation*. The increase in distance from the sources of the gravity anomalies diminishes them, effectively smoothing an anomaly map.

More often it is necessary to compute the gravity anomalies on a deeper surface than the measurement surface. This is called *downward continuation*. The inverse square law of gravitation results in amplification of the anomalies, the closer the new reference surface is to the source masses. This results in stronger, sharper anomalies, but it also exaggerates the higher-frequency noise in the data, and can lead to false interpretation.

Upward and downward continuation are also used in the interpretation of magnetic surveys, and in the analysis of the global magnetic field (Section 11.2.4).

4.2.3 Modeling Gravity Anomalies

After removal of regional effects, the residual gravity anomaly must be interpreted in terms of an anomalous density distribution. Modern analyses are based on iterative modeling using computers. Earlier methods of interpretation utilized comparison of the observed gravity anomalies with the computed anomalies of geometric shapes. The success of this simple approach is due to the insensitivity of the shape of a gravity anomaly to minor variations in the anomalous density distribution. Some fundamental problems of interpreting gravity anomalies can be learned from the computed effects of geometric models. In particular, it is important to realize that the interpretation of gravity anomalies is not unique; different density distributions can give the same anomaly.

4.2.3.1 Uniform Sphere: Model for a Diapir

Diapiric structures introduce material of different density into the host rock. A low-density salt dome ($\rho = 2150$ kg m^{-3}) intruding higher-density carbonate rocks ($\rho_0 = 2500$ kg m^{-3}) has a density contrast $\Delta\rho = -350$ kg m^{-3} and causes a negative gravity anomaly. A volcanic plug ($\rho = 2800$ kg m^{-3}) intruding a granite body ($\rho_0 = 2600$ kg m^{-3}) has a density contrast $\Delta\rho = +200$ kg m^{-3}, which causes a positive gravity anomaly. The contour lines on a map of the anomaly are centered on the diapir, so all profiles across the center of the structure are equivalent. The anomalous body can be modeled equally by a vertical cylinder or by a sphere, which we will evaluate here because of the simplicity of the model.

Assume a sphere of radius R and density contrast $\Delta\rho$ with its center at depth z below the surface (Fig. 4.19). The attraction Δg of the sphere is as though the anomalous mass M of the sphere were concentrated at its center. If we measure horizontal position from a point above its center, at distance x the vertical component Δg_z is given by

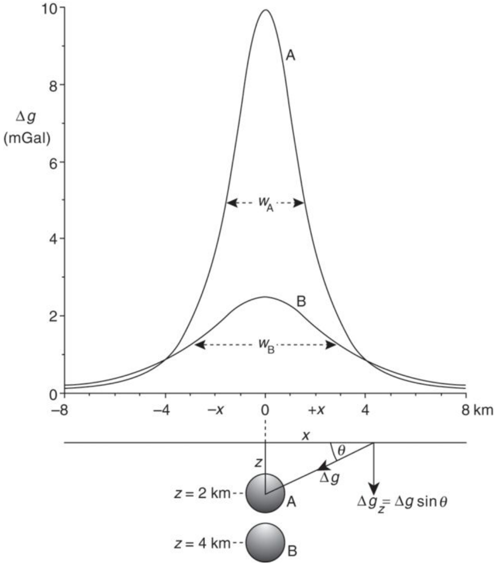


Fig. 4.19 Gravity anomalies for buried spheres with the same radius *R* and density contrast Δρ, but with their centers at different depths *z* below the surface. The anomaly of the deeper sphere B is flatter and broader than the anomaly of the shallower sphere A.

$$\Delta g_z = \Delta g \sin\theta = G\frac{M}{r^2}\frac{z}{r}, \tag{4.23}$$

where

$$M = \frac{4}{3}\pi R^3 \Delta\rho \quad \text{and} \quad r^2 = z^2 + x^2.$$

Substituting these expressions into Eq. (4.23) and rearranging terms gives

$$\Delta g_z = \frac{4}{3}\pi G \Delta\rho R^3 \frac{z}{(z^2+x^2)^{3/2}}$$
$$= \frac{4}{3}\pi G\left(\frac{\Delta\rho R^3}{z^2}\right)\left[\frac{1}{\left(1+(x/z)^2\right)}\right]^{3/2}.$$
(4.24)

The terms in the first pair of parentheses depend on the size, depth, and density contrast of the anomalous sphere. These properties determine the maximum amplitude of the anomaly, Δg_0, which is reached over the center of the sphere at $x = 0$. The peak value is given by

$$\Delta g_0 = \frac{4}{3}\pi G\left(\frac{\Delta\rho R^3}{z^2}\right).$$
(4.25)

This equation shows how the depth to the center of the sphere affects the peak amplitude of the anomaly; the greater the depth to the center, the smaller the amplitude (Fig. 4.19). For a given depth, the same peak anomaly can be produced by numerous combinations of $\Delta\rho$ and R; a large sphere with a low-density contrast can give an identical anomaly to a small sphere with a high-density contrast. The gravity data alone do not allow us to resolve this ambiguity.

The terms in the second pair of parentheses in Eq. (4.24) describe how the amplitude of the anomaly varies with distance along the profile. The anomaly is symmetrical with respect to x, reaches a maximum value Δg_0 over the center of the sphere ($x = 0$), and decreases to zero at great distances ($x = \infty$). Note that the larger the depth z, the more slowly the amplitude decreases laterally with increasing x. A deep source produces a smaller but broader anomaly than the same source at shallower depth. The width w of the anomaly where the amplitude has one-half its maximum value is called the "half-height width." The depth z to the center of the sphere is deduced from this anomaly width from the relationship $z = 0.652\,w$.

4.2.3.2 Horizontal Line Element

Many geologically interesting structures extend to great distances in one direction but have a cross-sectional shape that does not change significantly along the strike of the structure. If the length along strike were infinite, the two-dimensional variation of density in the area of a cross-section would suffice to model the structure. However, this is not really valid as the lateral extent is never infinite. As a general rule, if the length of the structure normal to the profile is more than 20 times its width or depth, it can be well approximated as being two-dimensional (2-D). Otherwise, the end effects due to the limited lateral extent of the structure must be taken into account in computing its anomaly. An elongate body that requires end corrections is sometimes referred to as a 2.5-D structure. For example, the mass distribution under elongated bodies like anticlines, synclines, and faults should be modeled as 2.5-D structures. Here, we will handle the simpler two-dimensional models of these structures.

Let an infinitely long linear mass distribution with mass m per unit length extend horizontally along the y-axis at depth z (Fig. 4.20). The contribution $d(\Delta g_z)$ to the vertical gravity anomaly Δg_z at the point P due to a small element of length dy is

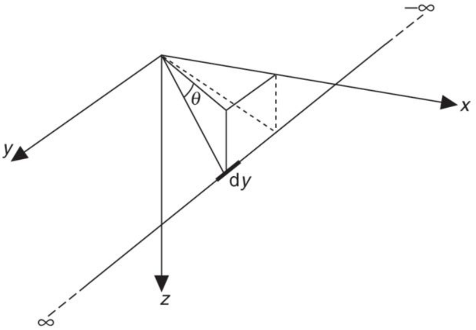


Fig. 4.20 Geometry for calculation of the gravity anomaly of an infinitely long linear mass distribution with mass m per unit length

extending horizontally along the y-axis at depth z.

$$d(\Delta g_z) = G\frac{m dy}{r^2}\sin\theta = G\frac{m dy}{r^2}\frac{z}{r}. \tag{4.26}$$

The line element extends from y = +∞ to y = −∞, so its vertical gravity anomaly is found by integration:

$$\Delta g_z = Gmz\int_{-\infty}^{\infty}\frac{dy}{r^3} = Gmz\int_{-\infty}^{\infty}\frac{dy}{(u^2+y^2)^{3/2}}, \tag{4.27}$$

where $u^2 = x^2 + z^2$. The integration is simplified by changing variables, so that $y = u \tan \varphi$; then dy $= u \sec^2\varphi$ dφ and $(u^2 + y^2)^{3/2} = u^3 \sec^3\varphi$. This gives

$$\Delta g_z = \frac{Gmz}{u^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\cos\varphi d\varphi, \tag{4.28}$$

which, after evaluation of the integral, gives

$$\Delta g_z = \frac{2Gmz}{z^2+x^2}. \tag{4.29}$$

This expression can be written as the derivative of a potential function ψ:

$$\Delta g_z = Gm\frac{2z}{u^2} = -\frac{\partial\psi}{\partial z}, \tag{4.30}$$

where

$$\Psi = Gm\log_e\left(\frac{1}{u}\right) = Gm\log_e\left(\frac{1}{\sqrt{x^2+z^2}}\right). \tag{4.31}$$

Ψ is called the *logarithmic potential*. Equations (4.29) and (4.31) are useful results for deriving formulae for the gravity anomaly of linear structures like an anticline (or syncline) or a fault.

4.2.3.3 Horizontal Cylinder: Model for Anticline or Syncline

The gravity anomaly of an anticline can be modeled by assuming that the upward folding of strata brings rocks with higher density nearer to the surface (Fig. 4.21a), thereby causing a positive density contrast. A syncline is modeled by assuming that its core is filled with strata of lower density that cause a negative density contrast. In each case the geometric model of the structure is an infinite horizontal cylinder (Fig. 4.21b).

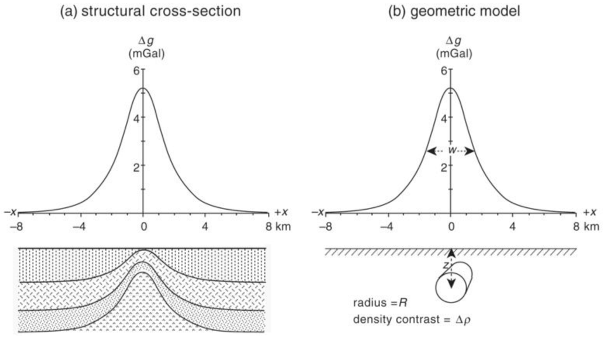


Fig. 4.21 Calculation of the gravity anomaly of an anticline: (a) structural cross-section, and (b) geometric model by an infinite horizontal cylinder.

A horizontal cylinder may be regarded as composed of numerous line elements parallel to its axis. The cross-sectional area of an element (Fig. 4.22) gives a mass anomaly per unit length, $m = \Delta\rho\, r\, d\theta\, dr$. The contribution $d\Psi$ of a line element to the potential at the surface is

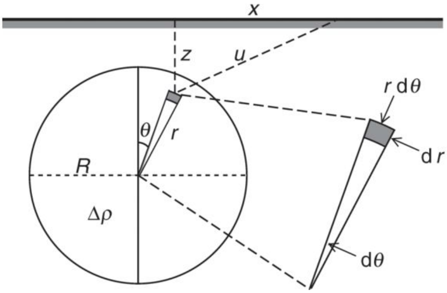


Fig. 4.22 Cross-sectional geometry for calculating the gravity anomaly of a buried horizontal cylinder made up of line elements parallel to its axis.

$$d\Psi = 2G\Delta\rho \log_e\left(\frac{1}{u}\right) r\,dr\,d\theta. \tag{4.32}$$

Integrating over the cross-section of the cylinder gives its potential Ψ; the vertical gravity anomaly of the cylinder is then found by differentiating Ψ with respect to z. Noting that $du/dz = z/u$, we get

$$\Delta g_z = -\frac{\partial\Psi}{\partial z} = -\frac{z}{u}\frac{\partial\Psi}{\partial u} = -\frac{2G\Delta\rho}{u}\int_0^{2\pi}\int_0^R \frac{\partial}{\partial u}\log_e\left(\frac{1}{u}\right) r\,dr\,d\theta. \tag{4.33}$$

After first carrying out the differentiation within the integral, this simplifies to

$$\Delta g_z = \frac{2G\Delta\rho z}{u^2}\int_0^{2\pi}\int_0^R r\,dr\,d\theta = \frac{2\pi G R^2\Delta\rho z}{x^2+z^2}. \tag{4.34}$$

Comparing Eqs. (4.34) and (4.29), it is evident that the gravity anomaly of the cylinder is the same as that of a linear mass element concentrated along its axis, with mass $m = \pi R^2 \Delta\rho$ per unit length along the strike of the structure. The anomaly can be written

$$\Delta g_z = 2\pi G\left(\frac{\Delta\rho R^2}{z}\right)\left[\frac{1}{1+(x/z)^2}\right].$$

(4.35)

The central peak value Δg_0 is given by

$$\Delta g_0 = 2\pi G\left(\frac{\Delta\rho R^2}{z}\right).$$

(4.36)

The shape of the anomaly on a profile normal to the structure (Fig. 4.21b) resembles that of a sphere (Fig. 4.19). However, the anomaly of a horizontal cylinder decreases laterally less rapidly than that of a sphere, due to the long extent of the cylinder normal to the profile. The “half-height width” w of the anomaly is again dependent on the depth z to the axis of the cylinder; in this case the depth is given by $z = 0.5\,w$.

4.2.3.4 Horizontal Thin Sheet

We next compute the anomaly of a thin horizontal ribbon of infinite length normal to the plane of the profile. This is done by fictively replacing the ribbon with numerous infinitely long line elements laid side by side. Let the depth of the sheet be z, its (infinitesimal) thickness t, and its density contrast $\Delta\rho$ (Fig. 4.23a); the mass per unit length in the y-direction of a line element of width dx is ($\Delta\rho\ t\,\mathrm{d}x$). Substituting in Eq. (4.29) gives the gravity anomaly of the line element; the anomaly of the thin ribbon is then computed by integrating between the limits x_1 and x_2 (Fig. 4.23b):

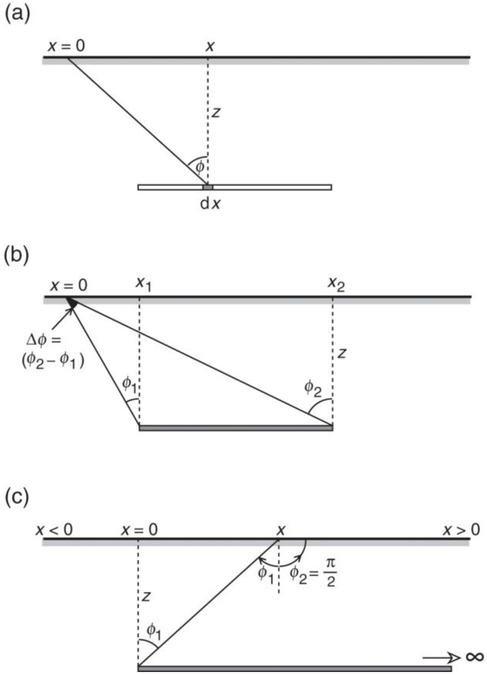


Fig. 4.23 Geometry for computation of the gravity anomaly across a horizontal thin sheet: (a) subdivision of the ribbon into line elements of width dx; (b) thin ribbon between the horizontal limits x_1 and x_2; and (c) semi-infinite horizontal thin sheet.

$$\Delta g_z = 2G\Delta\rho t z \int_{x_1}^{x_2} \frac{dx}{x^2 + z^2}$$

$$= 2G\Delta\rho t \left[\tan^{-1}\left(\frac{x_2}{z}\right) - \tan^{-1}\left(\frac{z_1}{z}\right)\right]. \tag{4.37}$$

Writing $\tan^{-1}(x_1/z) = \phi_1$ and $\tan^{-1}(x_2/z) = \phi_2$ as in Fig. 4.23b, the equation becomes

$$\Delta g_z = 2G\Delta\rho t[\phi_2 - \phi_1], \tag{4.38}$$

i.e., the gravity anomaly of the horizontal ribbon is proportional to the angle it subtends at the point of measurement.

The anomaly of a semi-infinite horizontal sheet is a limiting case of this result. For easier reference, the origin of x is moved to the edge of the sheet, so that distances to the left are negative and those to the right are positive (Fig. 4.23c). This makes $\phi_1 = -\tan^{-1}(x/z)$. The remote end of the sheet is at infinity, and $\phi_2 = \pi/2$. The gravity anomaly is then

$$\Delta g_z = 2G\Delta\rho t\left[\frac{\pi}{2} + \tan^{-1}\left(\frac{x}{z}\right)\right]. \tag{4.39}$$

A further example is the infinite horizontal sheet, which extends to infinity in the positive and negative x and y directions. With $\phi_2 = \pi/2$ and $\phi_1 = -\pi/2$ the anomaly is

$$\Delta g_z = 2\pi G\Delta\,\rho t, \tag{4.40}$$

which is the same as the expression for the Bouguer plate correction (Eq. 4.14).

4.2.3.5 Horizontal Slab: Model for a Vertical Fault

The gravity anomaly across a vertical fault increases progressively to a maximum value over the uplifted side (Fig. 4.24a). This is interpreted as due to the upward displacement of denser material, which causes a horizontal density contrast across a vertical step of height h (Fig. 4.24b). The faulted block can be modeled as a semi-infinite horizontal slab of height h and density contrast $\Delta\rho$, with its mid-point at depth z_0 (Fig. 4.24c).

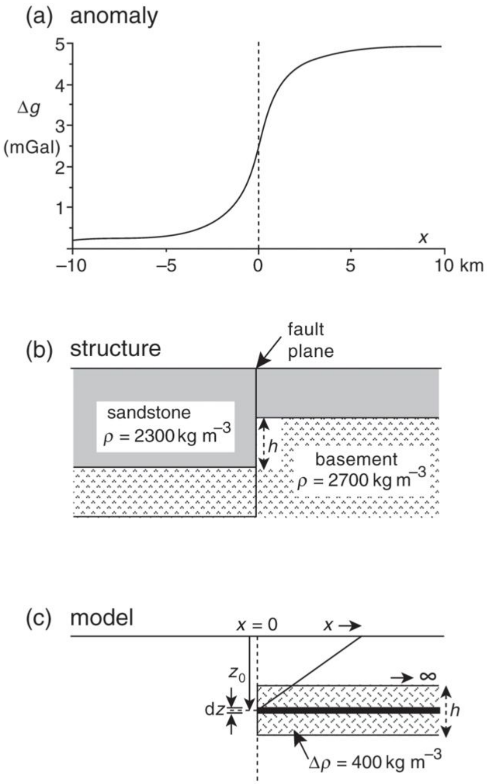

(a) anomaly
Δg
(mGal)
5
4
3
2
1
–10
–5
0
5
10 km
x
(b) structure
fault
plane
sandstone
ρ = 2300 kg m⁻³
h
basement
ρ = 2700 kg m⁻³
(c) model
x = 0
x →
z0
→ ∞
dz
h
Δρ = 400 kg m⁻³

Fig. 4.24 (a) The gravity anomaly across a vertical fault; (b) structure of a fault with vertical displacement *h*; and (c) model of the anomalous body as a semi-infinite horizontal slab of height *h*.

Let the slab be divided into thin, semi-infinite horizontal sheets of thickness dz at depth z. The gravity anomaly of a given sheet is given by Eq. (4.39) with dz for the thickness t. The anomaly of the semi-infinite slab is found by integrating with respect to z over the thickness of the slab; the limits of integration are $z-(h/2)$ and $z+(h/2)$. After slightly rearranging terms, this gives

$$\Delta g_z = 2G\Delta\,\rho h\left[\frac{\pi}{2}+\frac{1}{h}\int_{z_0-h/2}^{z_0+h/2}\tan^{-1}\left(\frac{x}{z}\right)dz\right]. \tag{4.41}$$

The second expression in the brackets is the mean value of the angle

$\tan^{-1}(x/z)$ averaged over the height of the fault step. This can be replaced to a good approximation by the value at the mid-point of the step, at depth z_0. This gives

$$\Delta g_z = 2G\Delta \rho h\left[\frac{\pi}{2} + \tan^{-1}\left(\frac{x}{z_0}\right)\right]. \tag{4.42}$$

Comparison of this expression with Eq. (4.39) shows that the anomaly of the vertical fault (or a semi-infinite thick horizontal slab) is the same as if the anomalous slab were replaced by a thin sheet of thickness h at the mid-point of the vertical step. Equation (4.42) is called the "thin-sheet approximation." It is accurate to about 2% provided that $z_0 > 2h$.

4.2.3.6 Iterative Modeling

The simple geometric models used to compute the gravity anomalies in the previous sections are crude representations of the real anomalous bodies. Modern computer algorithms have radically changed modeling methods by facilitating the use of an *iterative* procedure. A starting model with an assumed geometry and density contrast is postulated for the anomalous body. The gravity anomaly of the body is then computed and compared with the residual anomaly. The parameters of the model are changed slightly and the computation is repeated until the discrepancies between the model anomaly and the residual anomaly are smaller than a determined value. However, as in the case of the simple models, this process still does not give a unique solution for the density distribution.

Two- and three-dimensional iterative techniques are in widespread use. The two-dimensional method assumes that the anomalous body is infinitely long parallel to the strike of the structure, but end corrections for the possibly limited horizontal extent of the body can be made. We can imagine that the cross-sectional shape of the body is replaced by countless thin rods or line elements aligned parallel to the strike. Each rod makes a contribution to the vertical component of gravity at the origin (Fig. 4.25a). The gravity anomaly of the structure is calculated by adding up the contributions of all the line elements; mathematically, this is an integration over the end surface of the body. Although the theory is beyond the scope of this chapter, the gravity anomaly has the following simple form:

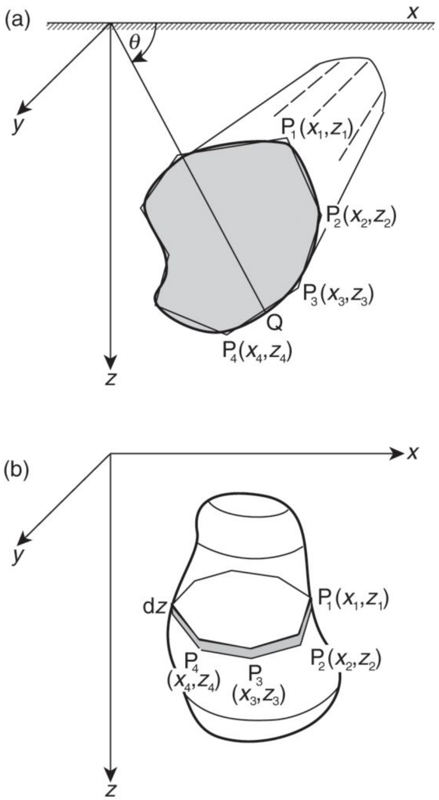


Fig. 4.25 Methods of computing gravity anomalies of irregular bodies: (a) the cross-section of a two-dimensional structure can be replaced with a multi-sided polygon. Source: M. Talwani, J.L. Worzel, and M. Landisman, Rapid gravity computations for two-dimensional bodies with application to the Mendocino submarine fracture zone, *J. Geophys. Res.* 64, 49–59, 1959. (b) a 3-D body can be replaced with thin horizontal laminae.

Source: M. Talwani and M. Ewing, Rapid computation of gravitational attraction of three-dimensional bodies of arbitrary shape, *Geophysics* 25, 203–225, 1960.

$$\Delta g_z = 2G\Delta\rho\oint z\,d\theta. \tag{4.43}$$

The angle θ is defined to lie between the positive x-axis and the radius from the origin to a line element (Fig. 4.25a), and the integration over the end-surface has been changed to an integration around its boundary. The computer algorithm for the calculation of this integral is greatly speeded up by replacing the true cross-sectional shape with an N-sided polygon (Fig. 4.25a). Apart from the assumed density contrast, the only important parameters for the computation are the (x, z) coordinates of the corners of the polygon. The origin is now moved to the next point on a profile across the structure. This move changes only the x-coordinates of the corners of the polygon. The calculations are repeated for each successive point on the profile. Finally, the calculated gravity anomaly profile across the structure is compared to the observed anomaly and the residual differences are evaluated. The coordinates of the corners of the polygon are adjusted and the calculation is reiterated until the residuals are less than a selected tolerance level.

The gravity anomaly of a 3-D body is modeled in a similar way. Suppose that we have a contour map of the body; the contour lines show the smooth outline of the body at different depths. We could construct a close replica of the body by replacing the material between successive contour lines with thin laminae. Each lamina has the same outline as the contour line and has a thickness equal to the contour separation. As a further approximation the smooth outline of each lamina is replaced by a multi-sided polygon (Fig. 4.25b). The gravity anomaly of the polygon at the origin is computed as in the 2-D case, using the (x, y) coordinates of the corners, the thickness of the lamina and an assumed density contrast. The gravity anomaly of the 3-D body at the origin is found by adding up the contributions of all the laminae. As in the simpler 2-D example, the origin is now displaced to a new point and the computation is repeated. The calculated and observed anomalies are now compared and the coordinates of the corners of the laminae are adjusted accordingly; the assumed density distribution can also be adjusted. The iterative procedure is repeated until the desired match between computed and observed anomalies is obtained.

4.2.4 Interpretation of Regional Gravity Anomalies

Without auxiliary information the interpretation of gravity anomalies is ambiguous, because the same anomaly can be produced by different bodies. An independent data source is needed to restrict the choices of density contrast, size, shape, and depths in the many possible gravity models. The additional information may be in the form of surface geological observations, from which the continuation of structures at depth is interpreted. Seismic refraction or reflection data provide better constraints.

The combination of seismic refraction experiments with precise gravity measurements has a long, successful history in the development of models of crustal structure. Refraction seismic profiles parallel to the trend of elongate geological structures give reliable information about the vertical velocity distribution. However, refraction profiles normal to the structural trend give uncertain information about the tilt of layers or lateral velocity changes. Lateral changes of crustal structure can be interpreted from several refraction profiles more or less parallel to the structural trend or from seismic reflection data. The refraction results give layer velocities and the depths to refracting interfaces. To compute the gravity effect of a structure, the velocity distribution must first be converted to a density model using a P-wave velocity–density relationship like the curve shown in Fig. 4.9. The theoretical gravity anomaly over the structure is computed using a 2-D or 3-D method. Comparison with the observed gravity anomaly (e.g., by calculating the residual differences point by point) indicates the plausibility of the model. It is always important to keep in mind that, because of the non-uniqueness of gravity modeling, a plausible structure is not necessarily the true structure. Despite the ambiguities some characteristic features of gravity anomalies have been established for important regions of the Earth.

4.2.4.1 Continental and Oceanic Gravity Anomalies

In examining the shape of the Earth we saw that the ideal reference figure is a spheroid, or ellipsoid of rotation. It is assumed that the reference Earth is in hydrostatic equilibrium. This is supported by observations of free-air and isostatic anomalies (Section 4.3.4), which suggest that, except in some unusual locations such as deep-sea trenches and island arcs in subduction zones, the continents and oceans are in approximate isostatic equilibrium with each other. By applying the concepts of isostasy (Section 4.3) we can understand the large-scale differences between Bouguer gravity anomalies over the continents and those over the oceans. In general, Bouguer anomalies over the continents are negative, especially over mountain ranges where the crust is unusually thick; in contrast, strongly positive Bouguer anomalies are found over oceanic regions where the crust is very thin.

The inverse relationship between Bouguer anomaly amplitude and crustal thickness can be explained with the aid of a hypothetical example (Fig. 4.26). Continental crust that has not been thickened or thinned by tectonic processes is considered to be “normal” crust. It is typically 30–35 km thick. Under location A on an undeformed continental coastal region a thickness of 34 km is assumed. The theoretical gravity used in computing a gravity anomaly is defined on the reference ellipsoid, the surface of which corresponds to mean sea level. Thus, at coastal location A on normally thick continental crust the Bouguer anomaly is close to zero. Isostatic compensation of the mountain range gives it a root-zone that increases the crustal thickness at location B. Seismic evidence shows that continental crustal density increases with depth from about 2700 kg m^{-3} in the upper granitic crust to about 2900 kg m^{-3} in the lower gabbroic crust. Thus, the density in the root-zone is much lower than the typical mantle density of 3300–3400 kg m^{-3} at the same depth under A. The low-density root beneath B causes a negative Bouguer anomaly, which typically reaches – 150 to –200 mGal.

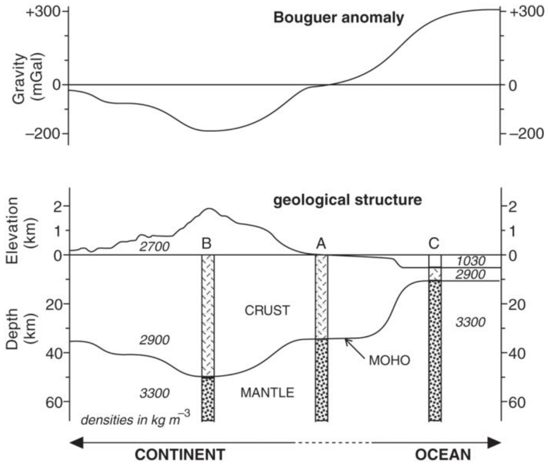

Bouguer anomaly
Gravity (mGal)
+300
0
−200
+300
0
−200
geological structure
Elevation (km)
Depth (km)
2
1
0
20
40
60
2700
B
A
C
1030
2900
CRUST
3300
2900
MOHO
3300
MANTLE
densities in kg m−3
CONTINENT
OCEAN

Fig. 4.26 Hypothetical Bouguer anomalies over continental and oceanic areas. The regional Bouguer anomaly varies roughly inversely with crustal thickness and topographic elevation.

Source: E.S. Robinson and C. Çoruh, *Basic Exploration Geophysics*, 562 pp., New York: John Wiley, 1988.

At oceanic location C the vertical crustal structure is very different. Two effects contribute to the Bouguer anomaly. A 5 km thick layer of seawater (density 1030 kg m^{-3}) overlies the thin basic oceanic crust (density 2900 kg m^{-3}), which has an average thickness of only about 6 km. To compute the Bouguer anomaly the seawater must be replaced by oceanic crustal rock. The attraction of the water layer is inherent in the measured gravity, so the density used in correcting for the Bouguer plate and the topography of the ocean bottom is the reduced density of the oceanic crust (i.e., 2900 – 1030 = 1870 kg m^{-3}). However, a more important effect is that the top of the mantle is at a depth of only 11 km. In a vertical section below this depth the mantle has a density of 3300–3400 kg m^{-3}, much higher than the density of the continental crust at equivalent depths below coastal site A. The lower 23 km of the section beneath C represents a large excess of mass. This gives rise to a strong positive Bouguer anomaly, which can amount to 300–400 mGal.

4.2.4.2 Gravity Anomalies Across Mountain Chains

The typical gravity anomaly across a mountain chain is strongly negative due to the large low-density root-zone. The Swiss Alps provide a good example of the interpretation of such a gravity anomaly with the aid of seismic refraction and reflection results. A precise gravity survey of Switzerland carried out in the 1970s yielded an accurate Bouguer gravity map (Fig. 4.27). The map contains effects specific to the Alps. Most obviously, the contour lines are parallel to the trend of the mountain range. In the south the negative anomaly displays a short-wavelength feature due to the northern extension of the positive anomaly of the so-called Ivrea Body. This is a high-density wedge of mantle material that was forced into an uplifted position within the western Alpine crust during an earlier continental collision. In addition, the Swiss gravity map contains the effects of low-density sediments that fill the Molasse Basin north of the Alps, the Po Plain to the south, and the major Alpine valleys.

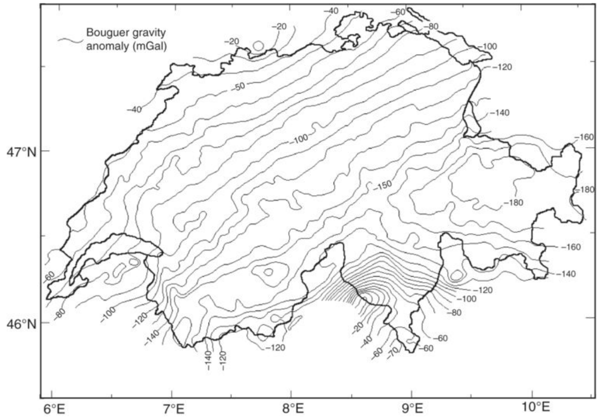
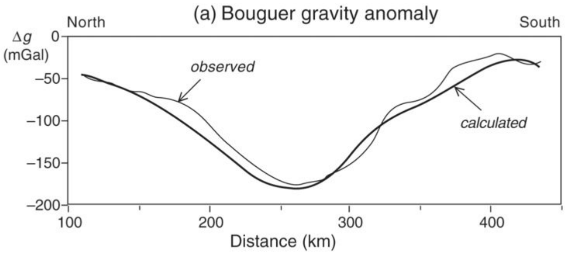


Fig. 4.27 Bouguer gravity map of Switzerland.

Source: E. Klingelé and R. Olivier, Die neue Schwere-Karte der Schweiz (Bouguer-Anomalien), Beitr. Geologie der Schweiz, *Serie Geophys*. 20, 93p, 1980.

In the late 1980s a coordinated geological and geophysical study – the European Geotraverse (EGT) – was made along and adjacent to a narrow path stretching from northern Scandinavia to northern Africa. Detailed reflection seismic profiles along its transect through the Central Swiss Alps complemented a large amount of new and extant refraction data. The seismic results gave the depths to important interfaces. Assuming a velocity–density relationship, a model of the density distribution in the lithosphere under the traverse was obtained. Making appropriate corrections for end effects due to limited extent along strike, a 2.5-D gravity anomaly was calculated for this lithospheric structure (Fig. 4.28). After removal of the effects of the high-density Ivrea Body and the low-density sediments, the corrected Bouguer gravity profile is reproduced well by the anomaly of the lithospheric model. Using the geometric constraints provided by the seismic data, the lithospheric gravity model favors a subduction zone, dipping gently to the south, with a high-density wedge of deformed lower crustal rock ("mélange") in the middle crust. As already noted, the fact that a density model delivers a gravity anomaly that agrees well with observation does not establish the reality of the interpreted structure, which can only be confirmed by further seismic imaging. However, the gravity model provides an important check on the reasonableness of suggested models. A model of crustal or lithospheric structure that does not give an appropriate gravity anomaly can reasonably be excluded.


(a) Bouguer gravity anomaly

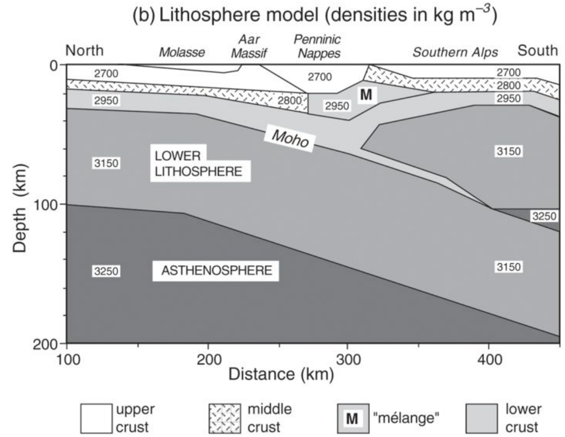

(b) Lithosphere model (densities in kg m^{-3})
North
Molasse
Aar Massif
Penninic Nappes
Southern Alps
South
0
100
200
Depth (km)
2700
2950
2800
2950
M
2700
2800
2950
Moho
3150
LOWER LITHOSPHERE
3150
3250
3250
ASTHENOSPHERE
3150
100
200
300
400
Distance (km)
upper crust
middle crust
M "mélange"
lower crust

Fig. 4.28 Lithosphere density model for the Central Swiss Alps along the European Geotraverse transect, compiled from seismic refraction and reflection profiles. The 2.5-D gravity anomaly calculated for this lithospheric structure is compared to the observed Bouguer anomaly after removal of the effects of the high-density Ivrea Body and the low-density sediments in the Molasse Basin, Po Plain, and larger Alpine valleys.

Source: K. Holliger and E. Kissling, Gravity interpretation of a unified 2-D acoustic image of the central Alpine collision zone, *Geophys. J. Int.* 111, 213–225, 1992.

4.2.4.3 Gravity Anomalies Across an Oceanic Ridge

An oceanic ridge system is a gigantic submarine mountain range. The difference in depth between the ridge crest and adjacent ocean basin is about 3 km. The ridge system extends laterally for several hundred kilometers on either side of the axis. Gravity and seismic surveys have been carried out across several oceanic ridge systems. Some common characteristics of continuous gravity profiles across a ridge are evident on a WNW–ESE transect crossing the Mid-Atlantic Ridge at 32°N (Fig. 4.29). The free-air gravity anomalies are small, around 50 mGal or less, and correlate closely with the variations in ocean-bottom topography. This indicates that the ridge and its flanks are nearly compensated isostatically. As expected for an oceanic profile, the Bouguer anomaly is strongly positive. It is greater than 350 mGal at distances beyond 1000 km from the ridge, but decreases to less than 200 mGal over the axis of the ridge.

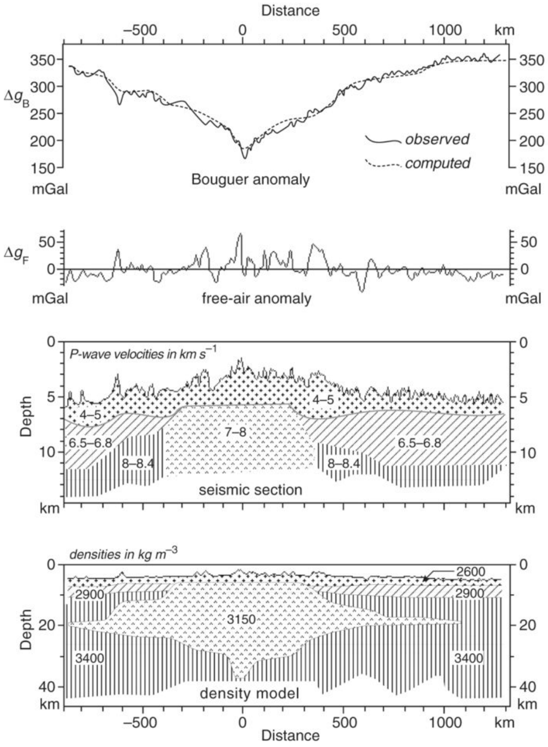


Fig. 4.29 Bouguer and free-air gravity anomalies over the Mid-Atlantic Ridge near 32°N. The seismic section is projected onto the gravity profile. The gravity anomaly computed from the density model fits the observed anomaly well, but is non-uniqu.

Source: M. Talwani, X. Le Pichon and M. Ewing, Crustal structure of the mid-ocean ridges. 2: Computed model from gravity and seismic refraction data, *J. Geophys. Res.* 70, 341–352, 1965.

The depths to important refracting interfaces and the P-wave layer velocities are known from seismic refraction profiles parallel to the ridge. The seismic structure away from the ridge is layered, with P-wave velocities of 4–5 km s^{-1} for the basalts and gabbros in oceanic Layer 2 and 6.5–6.8 km s^{-1} for the meta-basalts and meta-gabbros in Layer 3; the Moho is at about 11 km depth, below which typical upper mantle velocities of 8–8.4 km s^{-1} are found. However, the layered structure breaks down under the ridge at distances less than 400 km from its axial zone. Unusual velocities around 7.3 km s^{-1} occurred at several places, suggesting the presence of anomalous low-density mantle material at comparatively shallow depth beneath the ridge. The seismic structure was converted to a density model using a velocity–density relationship. Assuming a 2-D structure, several density models were found that closely reproduced the Bouguer anomaly. However, to satisfy the Bouguer anomaly on the ridge flanks each model requires a flat body in the upper mantle beneath the ridge; it extends down to about 30 km depth and for nearly 1000 km on each side of the axis (Fig. 4.29). The density of the anomalous structure is only 3150 kg m^{-3} instead of the usual 3400 kg m^{-3}. The model was proposed before the theory of plate tectonics was accepted. The anomalous upper mantle structure satisfies the gravity anomaly but has no relation to the known physical structure of a constructive plate margin. A broad zone of low upper mantle seismic velocities was not found in later experiments, but narrow low-velocity zones are sometimes present close to the ridge axis.

A further combined seismic and gravity study of the Mid-Atlantic Ridge near 46°N gave a contradictory density model. Seismic refraction results yielded P-wave velocities of 4.6 km s^{-1} and 6.6 km s^{-1} for Layers 2 and 3, respectively, but did not show anomalous mantle velocities beneath the ridge except under the median valley. A simpler density model was postulated to account for the gravity anomaly (Fig. 4.30). The small-scale free-air anomalies are accounted for by the variations in ridge topography seen in the bathymetric plot. The large-scale free-air gravity anomaly is reproduced well by a wedge-shaped structure that extends to 200 km depth. Its base extends to hundreds of kilometers on each side of the axis. A very small density contrast of only – 40 kg m^{-3} suffices to explain the broad free-air anomaly. The model is compatible with the thermal structure of an accreting plate margin. The low-density zone may be associated with hot material from the asthenosphere, which rises beneath the ridge, melts, and accumulates in a shallow magma chamber within the oceanic crust.

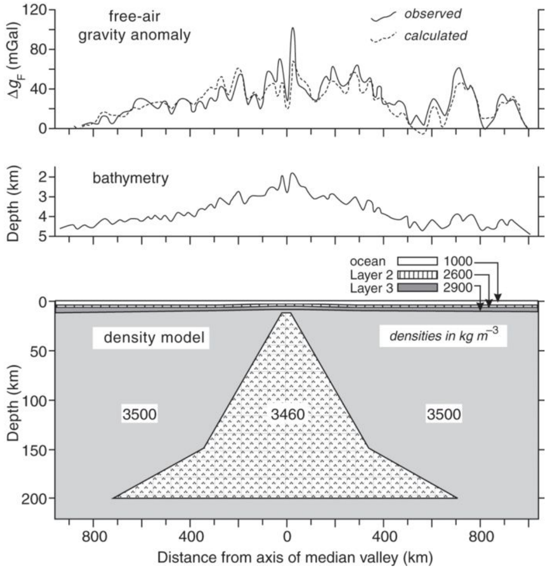


Fig. 4.30 Free-air gravity anomaly across the Mid-Atlantic Ridge near 46°N, and the lithospheric density model for the computed anomaly.

Source: C.E. Keen and C. Tramontini, A seismic refraction survey on the Mid-Atlantic Ridge, *Geophys. J. R. Astr. Soc.* 20, 473–491, 1970.

4.2.4.4 Gravity Anomalies at Subduction Zones

Subduction zones are found primarily at continental margins and island arcs. Elongate, narrow, and intense isostatic and free-air gravity anomalies have long been associated with island arcs. The relationship of gravity to the structure of a subduction zone is illustrated by the free-air anomaly across the Chile trench at 23°S (Fig. 4.31). Seismic refraction data define the thicknesses of the oceanic and continental crust. Thermal and petrological data are integrated to give a density model for the structure of the mantle and the subducting lithosphere.

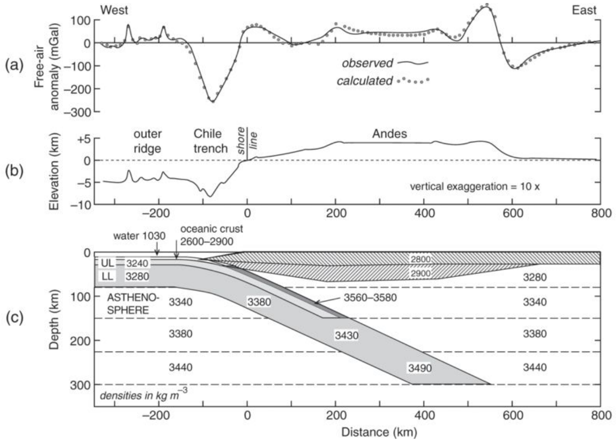


Fig. 4.31 Observed and computed free-air gravity anomalies across a subduction zone. The density model for the computed anomaly is based on seismic, thermal, and petrological data. The profile crosses the Chile trench and Andes mountains at 23°S.

Source: J.A. Grow and C.O. Bowin, Evidence for high-density crust and mantle beneath the Chile Trench due to the descending lithosphere, *J. Geophys. Res.* 80, 1449–1458, 1975.

The continental crust is about 65 km thick beneath the Andes mountains, and gives large negative Bouguer anomalies. The free-air gravity anomaly over the Andes is positive, averaging about +50 mGal over the 4 km high plateau. Even stronger anomalies up to +100 mGal are seen over the east and west boundaries of the Andes. This is largely due to the edge effect of the low-density Andean crustal block (see Fig. 4.13b and Section 4.1.6).

A strong positive free-air anomaly of about +70 mGal lies between the Andes and the shoreline of the Pacific Ocean. This anomaly is due to the subduction of the Nazca plate beneath South America. The descending slab is cool and dense, so there is a positive density contrast between the subducting lithosphere and the surrounding mantle. Also, petrological changes accompanying the subduction result in mass excesses. Peridotite in the upper lithosphere changes phase from plagioclase-type to the higher-density garnet-type. When oceanic crust is subducted to depths of 30–80 km, basalt changes phase to eclogite, which has a higher density (3560–3580 kg m^{-3}) than upper mantle rocks. These effects combine to produce the positive free-air anomaly.

The Chile trench is more than 2.5 km deeper than the ocean basin to the west. The sediments flooring the trench have low density. The mass deficiency of the water and sediments in the trench cause a strong negative free-air anomaly, which parallels the trench and has an amplitude greater than –250 mGal. A small positive anomaly of about +20 mGal is present about 100 km seaward of the trench axis. This anomaly is evident also in the mean level of the ocean surface as mapped by satellite altimetry measurements (Fig. 3.31), which shows that the mean sea surface is raised in front of deep ocean trenches. This is due to upward flexure of the lithosphere before its downward plunge into the subduction zone. The flexure elevates higher-density mantle rocks and thereby causes the small positive free-air anomaly.

4.3 Isostasy

4.3.1 The Discovery of Isostasy

Newton formulated the law of universal gravitation in 1687 and confirmed it with Kepler's laws of planetary motion. However, in the seventeenth and eighteenth centuries the law could not be used to calculate the mass or mean density of the Earth, because the value of the gravitational constant was not yet known (it was first determined by Cavendish in 1798). Meanwhile, eighteenth-century scientists attempted to estimate the mean density of the Earth by various means. These involved comparing the attraction of the Earth with that of a suitable mountain, which could be calculated. Inconsistent results were obtained.

During a French expedition to Peru and Ecuador in 1737–1740, Pierre Bouguer measured gravity with a pendulum at different altitudes, applying the correction term which now bears his name. If the density of crustal rocks is ρ and the mean density of the Earth is ρ_0, the ratio of the Bouguer plate correction (see Section 4.1.5.4) for elevation h to mean gravity for a spherical Earth of radius R is

$$\frac{\Delta g_{\mathrm{BP}}}{g} = \frac{2\pi G\rho h}{\frac{4}{3}\pi G\rho_0 R} = \frac{3}{2}\left(\frac{\rho}{\rho_0}\right)\left(\frac{h}{R}\right). \tag{4.44}$$

From the results he obtained near Quito, Ecuador, Bouguer estimated that the mean density of the Earth was about 4.5 times the density of crustal rocks.

The main method employed by Bouguer to determine the Earth's mean density consisted of measuring the deflection of the plumb-line (Section 3.2.3) by the mass of a nearby mountain (Fig. 4.32). Suppose the angle between the direction to a known star and the local vertical direction is measured at points N and S on the same meridian. The angles should be α_N and α_S, respectively. Their sum is α, the angle subtended at the center of the Earth by the radii to N and S, which corresponds to the difference in latitude. If N and S lie on opposite sides of a large mountain, the plumb-line at each station is deflected by the attraction of the mountain. As a result, the measured angles to the star are β_N and β_S, respectively, and their sum is β. The local vertical directions now intersect at point D instead of at the center of the (assumed spherical) Earth. The difference $\delta = \beta - \alpha$ is the sum of the deviations of the vertical direction caused by the mass of the mountain.

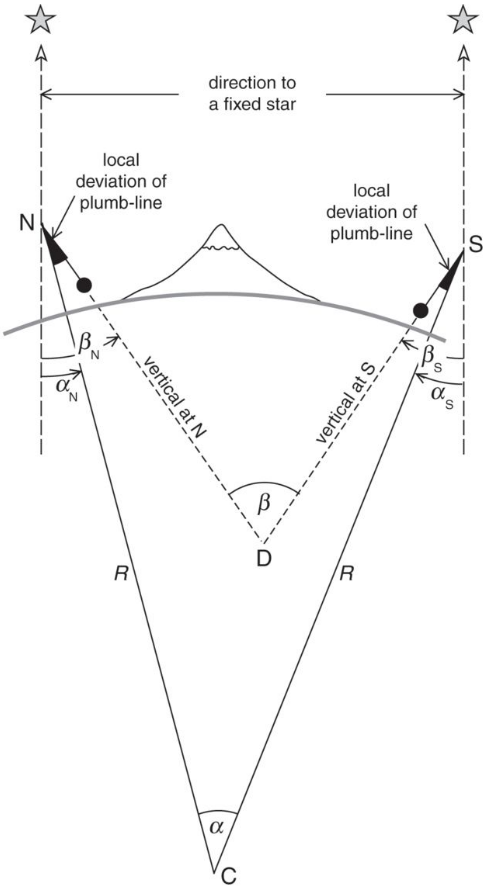

direction to a fixed star
local deviation of plumb-line
local deviation of plumb-line
N
S
β_N
α_N
β_S
α_S
vertical at N
vertical at S
β
D
R
R
α
C

Fig. 4.32 Deviations of the local plumb-line at positions N and S on opposite sides of a large mountain cause the local vertical directions to intersect at the point D instead of at the center of the Earth.

The horizontal attraction f of the mountain can be calculated from its shape and density, with a method that resembles the computation of the topographic correction in the reduction of gravity measurements. Dividing the mountain into vertical cylindrical elements, the horizontal attraction of each element is calculated and its component ($G\rho h_i$) toward the center of mass of the mountain is found. Summing up the effects of all the cylindrical elements in the mountain gives the horizontal attraction f toward its center of mass. Comparing f with mean gravity g, we then write

$$\tan\delta = \frac{f}{g} = \frac{G\rho \sum_i h_i}{\frac{4}{3}\pi G\rho_0 R} = \frac{3}{4}\frac{\sum_i h_i}{\pi R}\left(\frac{\rho}{\rho_0}\right). \tag{4.45}$$

For very small angles, $\tan\delta$ is equal to δ, and so the deflection of the vertical is proportional to the ratio ρ/ρ_0 of the mean densities of the mountain and the Earth. Bouguer measured the deflection of the vertical caused by Mt. Chimborazo (6272 m), Ecuador's highest mountain. His results suggested that the mean density of the Earth is around 12 times greater than the density of the mountain, which is unrealistically large and quite different from the values he had obtained near Quito. The erroneous result indicated that the deflection of the vertical caused by the mountain was too small for its estimated mass.

In 1774 Bouguer's Chimborazo experiment was repeated in Scotland by Neville Maskelyne on behalf of the Royal Society of London. Measurements of the elevations of stars were made on the north and south flanks of Mt. Schiehallion at sites that differed in latitude by 42.9″ of arc. The observed angle between the plumb-lines was 54.6″. The difference indicated that the Earth's density was 1.79 that of the mountain, which was estimated as 2500 kg m^{-3}. This result suggested a mean density for the Earth of 4500 kg m^{-3}, which was more realistic than Bouguer's unexplained result.

Further results accumulated in the first half of the nineteenth century. From 1806 to 1843 the English geodesist George Everest carried out triangulation surveys in India. He measured by triangulation the separation of a site at Kalianpur on the Indo–Ganges plain from a site at Kaliana in the foothills of the Himalayas. The distance differed substantially from the separation of the sites computed from the elevations of stars, as in Fig. 4.32. The discrepancy of 5.23″ of arc (162 m) was attributed to deflection of the plumb-line by the mass of the Himalayas. This would affect the astronomic determination but not the triangulation measurement. In 1855 J. H. Pratt computed the minimum deflection of the plumb-line that might be caused by the mass of the Himalayas and found that it should be 15.89″ of arc, about three times larger than the observed deflection. Evidently the attraction of the mountain range on the plumb-line was not as large as it should be.

The anomalous deflections of the vertical were first understood in the middle of the nineteenth century, when it was realized that there are regions beneath mountains – "root-zones" – in which rocks have a lower density than expected. The deflection of a plumb-line is not caused only by the horizontal attraction of the visible part of a mountain. The deficiency of mass at depth beneath the mountain means that the "hidden part" exerts a reduced lateral attraction, which partly offsets the effect of the mountain and diminishes the deflection of the vertical. In 1889 C. E. Dutton referred to the compensation of a topographic load by a less-dense subsurface structure as *isostasy*.

4.3.2 Models of Isostasy

Separate explanations of the anomalous plumb-line deflections were put forward by G. B. Airy in 1855 and J. H. Pratt in 1859. Airy was the Astronomer Royal and director of the Greenwich Observatory. Pratt was an archdeacon of the Anglican Church at Calcutta, India, and a devoted scientist. Their hypotheses have in common the compensation of the extra mass of a mountain above sea level by a less-dense region (or root) below sea level, but they differ in the way the compensation is achieved. The models were very successful and became widely used by geodesists, who developed them further. In 1909–1910, J. F. Hayford in the USA derived a mathematical model to describe the Pratt hypothesis. As a result, this theory of isostasy is often called the Pratt–Hayford scheme of compensation. Between 1924 and 1938 W. A. Heiskanen derived sets of tables for calculating isostatic corrections based on the Airy model. This concept of isostatic compensation has since been referred to as the Airy–Heiskanen scheme.

It became apparent that both models had serious deficiencies in situations that required compensation over a larger region. In 1931 F. A. Vening Meinesz, a Dutch geophysicist, proposed a third model in which the crust acts as an elastic plate. As in the other models, the crust floats buoyantly on a substratum, but its inherent rigidity spreads topographic loads over a broader region.

4.3.2.1 The Airy–Heiskanen Model

According to the Airy–Heiskanen model of isostatic compensation (Fig. 4.33a), an upper layer of the Earth “floats” on a denser fluid-like substratum, just as icebergs float in water. The upper layer is equated with the crust and the substratum with the mantle. The height of a mountain above sea level is much less than the thickness of the crust underneath it, just as the visible tip of an iceberg is much smaller than the subsurface part. The densities of the crust and mantle are assumed to be constant; the thickness of the root-zone varies in proportion to the elevation of the topography.

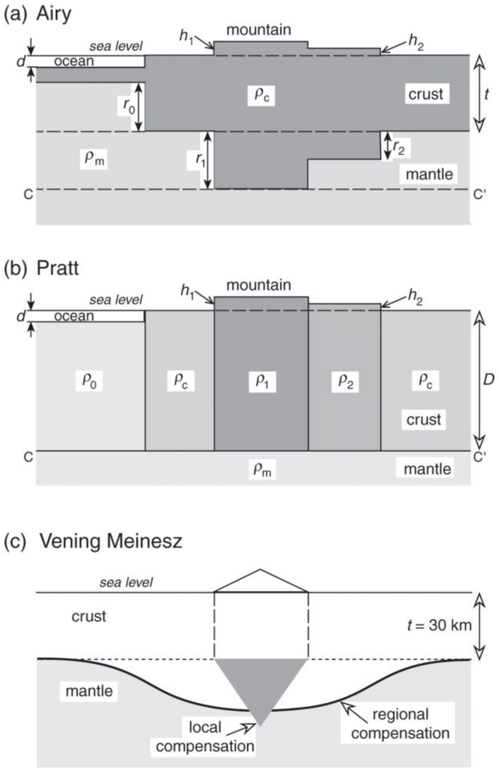

(a) Airy
sea level
ocean
d
h1
mountain
h2
ρc
crust
t
r0
r1
r2
ρm
mantle
C
C'
(b) Pratt
sea level
ocean
d
h1
mountain
h2
ρ0
ρc
ρ1
ρ2
ρc
crust
D
C
C'
ρm
mantle
(c) Vening Meinesz
sea level
crust
t = 30 km
mantle
local compensation
regional compensation

Fig. 4.33 Local isostatic compensation according to (a) the Airy–Heiskanen model and (b) the Pratt–Hayford model; (c) regional compensation according to the elastic plate model of Vening Meinesz.

The analogy to an iceberg is not exact, because under land at sea level the "normal" crust is already about 30–35 km thick; the compensating root-zone of a mountain lies below this depth. Oceanic crust is only about 10 km thick, thinner than the "normal" crust. The mantle between the base of the oceanic crust and the normal crustal depth is sometimes called the *anti-root* of the ocean basin.

The Airy–Heiskanen model assumes *local isostatic compensation*, i.e., the entire compensating mass of the root lies directly below each part of the excess topographic mass of the mountain. Isostasy is assumed to be complete, so that hydrostatic equilibrium exists at the compensation depth, which is equivalent to the base of the deepest mountain root. The pressure at this level is due to the weight of the rock material in the overlying vertical column (of basal area 1 m^2) extending to the Earth's surface. The vertical column for the mountain of height h_1 in Fig. 4.33a contains only crustal rocks of density ρ_c. The pressure at CC′ due to the mountain, "normal" crust of thickness t, and a root-zone of thickness r_1 amounts to $(h_1 + t + r_1)\rho_c$. The vertical column below the "normal" crust contains a thickness t of crustal rocks and thickness r_1 of mantle rocks; it exerts a pressure of $(t\rho_c + r_1\rho_m)$. For hydrostatic equilibrium the pressures are equal. Equating, and noting that each expression contains the term $t\rho_c$, we get

$$r_1 = \frac{\rho_c}{\rho_m - \rho_c} h_1, \tag{4.46}$$

with a similar expression for the root of depth r_2 under the hill of height h_2. The thickness r_0 of the anti-root of the oceanic crust under an ocean basin of water depth d and density ρw is given by

$$r_0 = \frac{\rho_c - \rho_w}{\rho_m - \rho_c} d. \tag{4.47}$$

The Airy–Heiskanen model assumes an upper layer of constant density floating on a more-dense substratum. It has root-zones of variable thickness proportional to the overlying topography. This scenario agrees broadly with seismic evidence for the thickness of the Earth's crust (see Section 7.2). The continental crust is much thicker than the oceanic crust. Its thickness is very variable, being largest below mountain chains, although the greatest thickness is not always under the highest topography. Airy-type compensation suggests hydrostatic balance between the crust and the mantle.

4.3.2.2 The Pratt–Hayford Model

The Pratt–Hayford isostatic model incorporates an outer layer of the Earth that rests on a weak magmatic substratum. Differential expansion of the material in vertical columns of the outer layer accounts for the surface topography, so that the higher the column above a common base the lower the mean density of rocks in it. The vertical columns have constant density from the surface to their base at depth D below sea level (Fig. 4.33b). If the rock beneath a mountain of height h_i ($i = 1, 2, \ldots$) has density ρ_i, the pressure at CC′ is $\rho_i(h_i + D)$. Beneath a continental region at sea level the pressure of the rock column of density ρ_c is $\rho_c D$. Under an ocean basin the pressure at CC′ is due to water of depth d and density $\rho_w$on top of a rock column of thickness $(D - d)$ and density ρ_0; it is equal to $\rho_w d + \rho_0(D - d)$. Equating these pressures, we get

$$\rho_i = \frac{D}{h_i + D}\rho_c \tag{4.48}$$

for the density below a topographic elevation h_i, and

$$\rho_0 = \frac{\rho_c D - \rho_w d}{D - d} \tag{4.49}$$

for the density under an oceanic basin of depth d. The compensation depth D is about 100 km.

The Pratt–Hayford and Airy–Heiskanen models represent *local isostatic compensation*, in which each column exerts an equal pressure at the compensation level. At the time these models were proposed very little was yet known about the internal structure of the Earth.

This was only deciphered after the development of seismology in the late nineteenth and early twentieth centuries. Each model is idealized, both with regard to the density distributions and the behavior of Earth materials. For example, the crust is assumed to offer no resistance to shear stresses arising from vertical adjustments between adjacent columns, yet it has sufficient strength to resist stresses due to horizontal differences in density. It is implausible that small topographic features require compensation at large depths; more likely, they are entirely supported by the strength of the Earth's crust.

4.3.2.3 Vening Meinesz Elastic Plate Model

In the 1920s, F. A. Vening Meinesz made extensive gravity surveys at sea. His measurements were made in a submarine to avoid the disturbances of wave motions. He studied the relationship between topography and gravity anomalies over prominent topographic features, such as the deep-sea trenches and island arcs in southeastern Asia, and concluded that isostatic compensation is often not entirely local. In 1931 he proposed a model of *regional isostatic compensation* which, like the Pratt–Hayford and Airy–Heiskanen models, envisages a light upper layer that floats on a denser fluid substratum. However, in the Vening Meinesz model the upper layer behaves like an elastic plate overlying a weak fluid. The strength of the plate distributes the load of a surface feature (e.g., an island or seamount) over a horizontal distance wider than the feature (Fig. 4.33c). The topographic load bends the plate downward into the fluid substratum, which is pushed aside. The buoyancy of the displaced fluid forces it upward, giving support to the bent plate at distances well away from the central depression. The bending of the plate, which accounts for the *regional* isostatic compensation in the Vening Meinesz model, depends on the elastic properties of the lithosphere.

4.3.3 Isostatic Compensation and Vertical Crustal Movements

In the Pratt–Hayford and Airy–Heiskanen models, the lighter crust floats freely on the denser mantle. The system is in hydrostatic equilibrium, and local isostatic compensation is a simple application of Archimedes' principle. A "normal" crustal thickness for sea-level coastal regions is assumed (usually 30–35 km) and the additional depths of the root-zones below this level are exactly proportional to the elevations of the topography above sea level. The topography is then *completely compensated* (Fig. 4.34a). However, isostatic compensation is often incomplete. The geodynamic imbalance leads to vertical crustal movements.

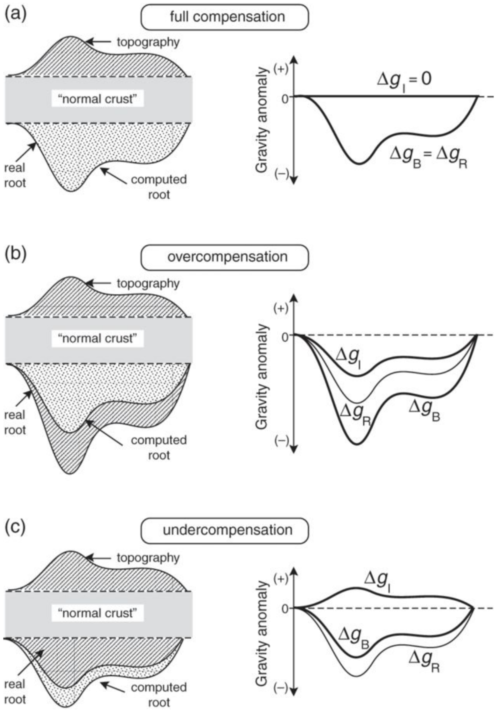

(a)
full compensation
topography
"normal crust"
real root
computed root
Gravity anomaly
(+)
0
(–)
$\Delta g_I = 0$
$\Delta g_B = \Delta g_R$
(b)
overcompensation
topography
"normal crust"
real root
computed root
Gravity anomaly
(+)
0
(–)
Δg_I
Δg_R
Δg_B
(c)
undercompensation
topography
"normal crust"
real root
computed root
Gravity anomaly
(+)
0
(–)
Δg_I
Δg_B
Δg_R

Fig. 4.34 Explanation of the isostatic gravity anomaly (Δg_I) as the difference between the Bouguer gravity anomaly (Δg_B) and the computed anomaly (Δg_R) of the root-zone estimated from the topography for (a) complete isostatic compensation, (b) isostatic overcompensation, and (c) isostatic undercompensation.

Mountains are subject to erosion, which can disturb isostatic compensation. If the eroded mountains are no longer high enough to justify their deep root-zones, the topography is isostatically *overcompensated* (Fig. 4.34b). Buoyancy forces are created, just as when a wooden block floating in water is pressed downward by a finger; the underwater part becomes too large in proportion to the amount above the surface. If the finger pressure is removed, the block rebounds in order to restore hydrostatic equilibrium. Similarly, the buoyancy forces that result from overcompensation of mountainous topography cause vertical *uplift*, and if this uplift is ongoing then the topography has not yet reached a state of *isostatic equilibrium.* The opposite scenario is also possible. When the visible topography has roots that are too small, the topography is isostatically *undercompensated* (Fig. 4.34c). This situation can result, for example, when tectonic forces thrust crustal blocks on top of each other. Ongoing *subsidence* of the uplifted region acts to bring the system back into isostatic equilibrium.

The most striking and best-observed examples of vertical crustal movements due to isostatic imbalance are related to the phenomenon of glacial rebound observed in northern Canada and in Fennoscandia. During the latest ice age these regions were covered by a thick ice-cap. The weight of ice depressed the underlying crust. Subsequent melting of the ice-cap removed the extra load on the crust, and it has since been rebounding. The post-glacial uplift of the Fennoscandian shield has been measured during the last 100 years using tide-gauge observations at coastal locations and repeated high-precision leveling surveys. These are augmented in modern analyses by continuous GNSS observations at a network of stations. As a result, the contour lines of equal uplift rate (Fig. 4.35) are well established even for inaccessible regions. The general pattern of glacial rebound is clearly recognizable, with uplift rates of up to 8 mm yr^{-1}.

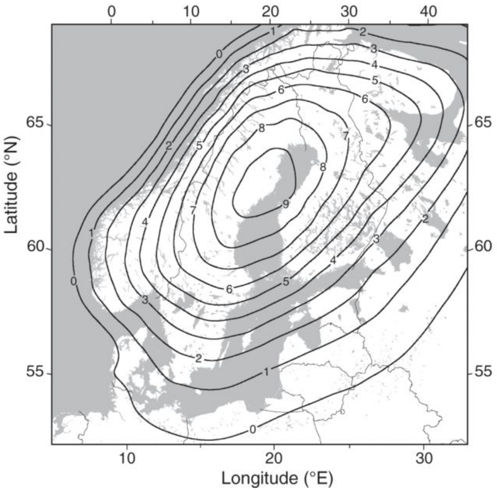

0
10
20
30
40
Latitude (°N)
65
60
55
10
20
30
Longitude (°E)
0
1
2
3
4
5
6
7
8
9

Fig. 4.35 Fennoscandian rates of vertical crustal movement (in mm yr^{-1}) relative to mean sea level, derived from repeated precise leveling, tide-gauges, and continuous GPS measurements. Positive rates correspond to uplift, negative rates to subsidence.

Source: Nordic Geodetic Commission model 2006. Courtesy of M. Poutanen.

4.3.4 Isostatic Gravity Anomalies

The different degrees of isostatic compensation find expression in gravity anomalies. As explained in Section 4.1.6, the free-air gravity anomaly Δg_F is small near the center of a large region that is isostatically compensated; the Bouguer anomaly Δg_B is strongly negative. Assuming complete isostatic compensation, the size and shape of the root-zone can be determined from the elevations of the topography. With a suitable density contrast the gravity anomaly Δg_r of the modeled root-zone can be calculated; because the root-zone has lower density than adjacent mantle rocks Δg_r is also negative. The *isostatic gravity anomaly* Δg_I is defined as the difference between the Bouguer gravity anomaly and the computed anomaly of the root-zone, i.e.,

$$\Delta g_I = \Delta g_B - \Delta g_r. \tag{4.50}$$

Examples of the isostatic gravity anomaly for the three types of isostatic compensation are shown schematically in Fig. 4.34. When isostatic compensation is complete, the topography is in hydrostatic equilibrium with its root-zone. Both Δg_B and Δg_r are negative but equal; consequently, the isostatic anomaly is everywhere zero ($\Delta g_I = 0$). In the case of overcompensation, the eroded topography suggests a root-zone that is smaller than the real root-zone. The Bouguer anomaly is caused by the larger real root, so Δg_B is numerically larger than Δg_r. Subtracting the smaller negative anomaly of the computed root-zone leaves a negative isostatic anomaly ($\Delta g_I < 0$). On the other hand, with undercompensation the topography suggests a root-zone that is larger than the real root-zone. The Bouguer anomaly is caused by the smaller real root, so Δg_B is numerically smaller than Δg_r. Subtracting the larger negative anomaly of the root-zone leaves a positive isostatic anomaly ($\Delta g_I > 0$).

A national gravity survey of Switzerland carried out in the 1970s gave a high-quality map of Bouguer gravity anomalies (see Fig. 4.27). Seismic data gave representative parameters for the Central European crust and mantle: a crustal thickness of 32 km without topography, and mean densities of 2670 kg m^{-3} for the topography, 2810 kg m^{-3} for the crust, and 3310 kg m^{-3} for the mantle. Using the Airy–Heiskanen model of compensation, a map of isostatic gravity anomalies in Switzerland was derived (Fig. 4.36) after correcting the gravity map for the effects of low-density sediments in the Molasse Basin north of the Alps and high-density material in the anomalous Ivrea Body in the south.

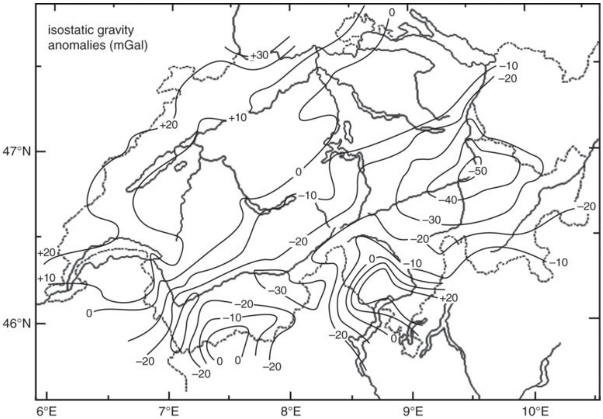


Fig. 4.36 Isostatic gravity anomalies in Switzerland, based on the national gravity map, corrected for the effects of the Molasse Basin and the Ivrea Body.

Source: E. Klingelé and E. Kissling, 1982. Zum Konzept der isostatischen Modelle in Gebirgen am Beispiel der Schweizer Alpen, Geodätisch-Geophysikalische Arbeiten in der Schweiz. *Schweiz. Geodät. Komm.*, 35, 3–36.

The pattern of isostatic anomalies reflects the different structures beneath the Jura mountains, which do not have a prominent root-zone, and the Alps, which have a low-density root that extends to more than 55 km depth in places. The dominant ENE–WSW trend of the isostatic gravity anomaly contour lines is roughly parallel to the trends of the mountain chains. In the northwest, near the Jura mountains, positive isostatic anomalies exceed 20 mGal. In the Alps, isostatic anomalies are mainly negative, reaching more than –50 mGal in the east.

A computation based on the Vening Meinesz model gave an almost identical isostatic anomaly map. The agreement of maps based on the different concepts of isostasy is somewhat surprising. It may imply that vertical crustal columns are not free to adjust relative to one another without friction, as assumed in the Airy–Heiskanen model. The friction is thought to result from horizontal compressive stresses in the Alps, which are active in the ongoing mountain-building process.

Comparison of the isostatic anomaly map with one of recent vertical crustal movements (Fig. 4.37) illustrates the relevance of isostatic gravity anomalies for tectonic interpretation. Precise leveling surveys have been carried out since the early 1900s along major valleys transecting and parallel to the mountainous topography of Switzerland. Relative rates of uplift or subsidence are computed from the differences between repeated surveys. The results have not been tied to absolute tide-gauge observations and so are relative to a base station at Aarburg in the canton of Aargau, in the northeast.

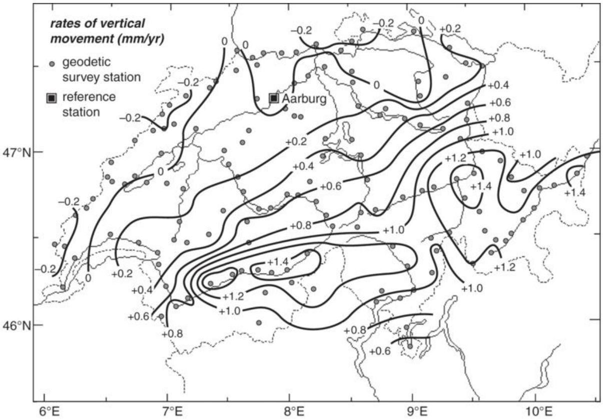


Fig. 4.37 Rates of vertical crustal motion in Switzerland deduced from repeated precise leveling. Broken contour lines indicate areas

in which geodetic data are absent or insufficient. Positive rates correspond to uplift, negative rates to subsidence (data source: Gubler, 1991).

The rates of relative vertical movement in northeastern Switzerland are smaller than the confidence limits on the data and may not be significant, but the general tendency suggests subsidence. This region is characterized by mainly positive isostatic anomalies. The rates of vertical movement in the southern part of Switzerland exceed the noise level of the measurements and are significant. The most notable characteristic of the recent crustal motions is vertical uplift of the Alpine part of Switzerland relative to the central plateau and Jura mountains. The Alpine uplift rates are up to 1.5 mm yr^{-1}, considerably smaller than the rates observed in Fennoscandia. The most rapid uplift rates are observed in the region where isostatic anomalies are negative. The constant erosion of the mountain topography relieves the crustal load and the isostatic response is uplift. However, the interpretation is complicated by the fact that compressive stresses throughout the Alpine region acting on deep-reaching faults can produce non-isostatic uplift of the surface. The separation of isostatic and non-isostatic vertical crustal movements in the Alps will require detailed and exact information about the structure of the lithosphere and asthenosphere in this region.

4.4 *Suggestions for Further Reading*

Introductory Level

Kearey, P., Brooks, M., and Hill, I. 2002. *An Introduction to Geophysical Exploration* (3rd ed.). Oxford: Blackwell Publishing.

Intermediate and Advanced Level

Blakely, R. J. 1995. *Potential Theory in Gravity and Magnetic Applications*. Cambridge: Cambridge University Press.

Dobrin, M. B. and Savit, C. H. 1988. *Introduction to Geophysical Prospecting* (4th ed.). New York: McGraw-Hill.

Reynolds, J. M. 2011. *An Introduction to Applied and Environmental Geophysics* (2nd ed.). Chichester: Wiley-Blackwell.

Telford, W. M., Geldart, L. P., and Sheriff, R. E. 1990. *Applied Geophysics*. Cambridge: Cambridge University Press.

4.5 Review Questions

1. Describe the principle of operation of a gravimeter.

2. Explain why a gravimeter only gives relative measurements of gravity.

3. What is the topographic correction in the reduction of gravity data? Why is it needed?

4. What are the Bouguer plate and free-air gravity corrections?

5. What is a free-air gravity anomaly? How does it differ from a Bouguer anomaly?

6. Sketch how the Bouguer gravity anomaly might vary on a continuous profile that extends from a continental mountain range to an oceanic ridge.

7. Describe *two* borehole methods for determining the density of rocks around the borehole.

8. Describe and explain the Nettleton profile method to determine the optimum density for interpreting a gravity survey.

9. What is isostasy? What is an isostatic gravity anomaly?

10. Why is the free-air gravity anomaly close to zero at the middle of a large crustal block that is in isostatic equilibrium?

11. Describe the three models of isostasy, and explain how they differ from each other.

12. What would be the geodynamic behavior of a region that is characterized by a negative isostatic gravity anomaly?

4.6 Exercises

1. (a) By differentiating the normal gravity formula given by Eq. (3.59), derive an expression for the change of gravity with latitude. Calculate the gravity change in milligal per kilometer of northward displacement at latitudes 30°N, 45°N, and 60°N.

(b) Repeat the exercise using the International Gravity Formula, Eq. (3.62).

2. The following gravity measurements were made on a traverse across a rock formation. Use the combined elevation correction to compute the apparent density of the rock.

Elevation (m)	Gravity (mGal)
100	–39.2
150	–49.5
235	–65.6
300	–78.1
385	–95.0
430	–104.2

3. Show that the "half-width" w of the gravity anomaly over a sphere and the depth z to the center of the sphere are related by $z = 0.652\, w$.

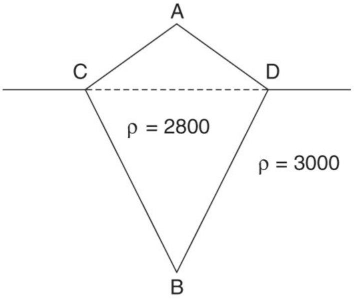
4. Assume the "thin-sheet approximation" (Eq. 4.42) for the gravity anomaly over a vertical fault of density contrast $\Delta\rho$ and height h with mid-point at depth z_0.

(a) What is the maximum slope of the anomaly and where does it occur?

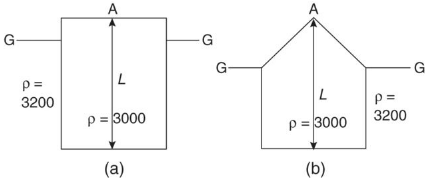
(b) Determine the relationship between the depth z_0 and the horizontal distance w between the positions where the slope of the anomaly is one-half the maximum slope.

5. Calculate the maximum gravity anomaly at ground level over a buried anticlinal structure, modeled by a horizontal cylinder with radius 1000 m and density contrast 200 kg m–^{3}, when the depth of the cylinder axis is (a) 1500 m and (b) 5000 m.

6. The peak A of a mountain is 1000 meters above the level CD of the surrounding plain, as in the diagram. The density of the rocks forming the mountain is 2800 kg m–^{3}, that of the surrounding crust is 3000 kg m–^{3}. Assuming that the mountain and its "root" are symmetric about AB and that the system is in local isostatic equilibrium, calculate the depth of B below the level CD.

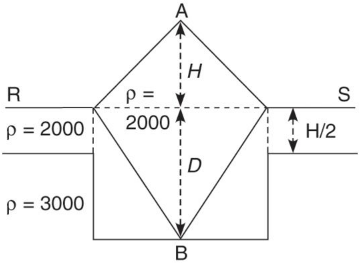


7. A crustal block with mean density 3000 kg m–^{3} is initially in isostatic equilibrium with the surrounding rocks whose density is 3200 kg m–^{3}, as in the figure (a). After subsequent erosion the above-surface topography is as shown in (b). The distance L remains constant (i.e., there is no erosion at the highest point A) and Airy-type isostatic equilibrium is maintained. Calculate in terms of L the amount by which the height of A is changed. Explain why A moves in the sense given by your answer.



8. An idealized mountain-and-root system, as in the figure, is in isostatic equilibrium. The densities in kg m–^{3} are as shown.

Express the height H of the point A above the horizontal surface RS in terms of the depth D of the root B below this surface.


A
H
R
ρ =
S
2000
ρ = 2000
H/2
D
ρ = 3000
B

4.7 Computational Exercises

Jupyter notebook computational exercises can be found online for this chapter at www.cambridge.org/FoG3ed ((1) Fourier series, (2) Gravity corrections).

5

Rheology of the Earth

Preview

The Earth is continuously being deformed due to forces related to earthquakes, surface loads such as ice-sheets, the gravitational attraction to other planets, and many other phenomena. The relation between forces and the resulting style of deformation defines the rheology of a material. While the Earth behaves nearly like elastic rubber at short timescales from seconds to years, large parts of it can be treated as a liquid when studied over geologic timescales exceeding millennia. In this chapter we provide a detailed introduction to the various deformation styles of the Earth and their relevance to a wide range of processes, including the propagation of seismic waves, the deformation of the lithosphere near mountain ranges and subduction zones, and convective flow in the mantle.

5.1 Elastic Deformation

5.1.1 Elastic Behavior of Materials

When a force is applied to a solid material, it deforms. This means that the particles of the material are displaced from their original positions. Provided the force does not exceed a critical value, the displacements may be (1) instantaneous, (2) directly proportional to the force, and (3) reversible. In other words, the particles are displaced immediately upon application of the force, they are displaced twice as much when the force doubles in magnitude, and they return to their original positions when the force is removed, not leaving any permanent deformation. Materials that exhibit this kind of behavior are called *elastic*.

On the atomic level, elastic deformation only changes the relative positions of atoms in a solid without breaking chemical bonds between them. Elasticity is the simplest of all rheologies, and it will serve us to introduce basic concepts of deformation in response to forces.

The laws of elastic deformation are illustrated by the following example. Consider a right cylindrical block of height h and cross-sectional area A, subjected to a force F which acts to extend the block by the amount Δh (Fig. 5.1). Laboratory experiments on a wide range of materials, including rocks, show that Δh is directly proportional to the applied force and to the dimension of the block, but is inversely proportional to the cross-section of the block. That is, $\Delta h \propto Fh/A$, or

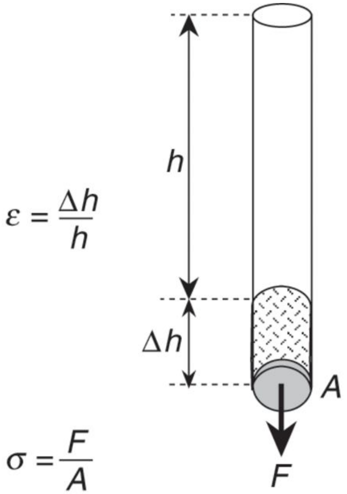


Fig. 5.1 A force *F* acting on a bar with cross-sectional area *A* extends the original length *h* by the amount Δ*h*. Hooke's law of elastic deformation states that Δ*h*/*h* is proportional to *F*/*A*.

$$\frac{F}{A} \propto \frac{\Delta h}{h}. \tag{5.1}$$

When the area *A* becomes infinitesimally small, the limiting value of the force per unit area (*F*/*A*) is called the stress, *σ*. The units of stress are the same as the units of pressure. The SI unit is the pascal, equivalent to a force of 1 newton per square meter (1 Pa = 1 N m^{-2}); the c.g.s. unit is the bar, equal to 10^6 dyne cm^{-2}, which is approximately equivalent to atmospheric pressure at sea level. Geological stresses and pressures in the Earth are measured in megabars or gigapascals (1 Mbar = 100 GPa).

When *h* is infinitesimally small, the fractional change in dimension (Δ*h*/*h*) is called the *strain* ε, which is a dimensionless quantity. Equation (5.1) states that, for elastic behavior, the strain in a body is proportional to the stress applied to it, i.e., *σ* and ε are related by a constant. This is referred to as *Hooke's law*.

When stresses are applied for long times, typically exceeding thousands of years, materials behave elastically only when pressure and temperature are low. This is the case, for example, within the Earth's crust, which deforms elastically even under the influence of long-term tectonic stresses, as long as they are not too large. At shorter timescales of seconds, minutes, or hours, the whole Earth, from crust to core, behaves approximately like an elastic solid. This is relevant for the propagation of seismic waves excited by earthquakes, explosions, or volcanic eruptions.

Elasticity is an idealization of material deformation under stress. While it is very useful for the description of a wide range of processes, Earth materials may in fact exhibit a more complex behavior. The material may show a somewhat delayed response to an applied stress. When the stress exceeds a certain threshold, the deformation may not fully recover when the stress is removed, and at even higher stresses, the material may fail. Furthermore, a solid, crystalline material may actually behave like a fluid over long timescales when pressure and temperature are sufficiently high. These deviations from ideal elasticity will be the topics of Sections 5.2 and 5.3.

We have, so far, considered stress and strain in only one direction. To generalize this concept to arbitrary directions (e.g., perpendicular to the rotation axis of the bar shown in Fig. 5.1), we now introduce the definitions of the stress and strain matrices.

5.1.2 The Stress Matrix

Consider a force **F** acting on a rectangular prism P in a reference frame defined by orthogonal Cartesian coordinate axes x, y, and z (Fig. 5.2a). The component of **F** which acts in the direction of the x-axis is designated F_x; the force **F** is fully defined by its components F_x, F_y, and F_z. The size of a small surface element is characterized by its area A, while its orientation is described by the direction normal to the surface (Fig. 5.2b). The small surface with area normal to the x-axis is designated A_x. The component of force F_x acting normal to the surface A_x produces a *normal stress*, denoted by σ_{xx}. The components of force along the y- and z-axes result in *shear stresses* σ_{yx} and σ_{zx} (Fig. 5.2c), given by

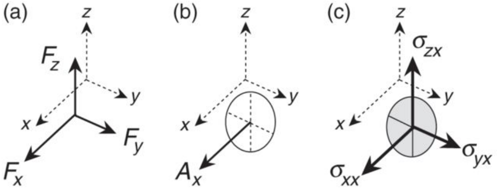


Fig. 5.2 (a) Components F_x, F_y, and F_z of the force **F** acting in a reference frame defined by orthogonal Cartesian coordinate axes x, y, and z. (b) The orientation of a small surface element with area A_x is described by the direction normal to the surface. (c) The components of force parallel to the x-axis result in the normal stress σ_{xx}; the components parallel to the y- and z-axes cause shear stresses σ_{xy} and σ_{xz}.

$$\sigma_{xx} = \lim_{A_x \to 0}\left(\frac{F_x}{A_x}\right) \quad \sigma_{yy} = \lim_{A_y \to 0}\left(\frac{F_y}{A_y}\right) \quad \sigma_{zz} = \lim_{A_z \to 0}\left(\frac{F_z}{A_z}\right). \tag{5.2}$$

Similarly, the components of the force **F** acting on an element of surface A_y normal to the y-axis define a normal stress σ_{yy} and shear stresses σ_{xy} and σ_{zy}, while the components of **F** acting on an element of surface A_z normal to the z-axis define a normal stress σ_{zz} and shear stresses σ_{xz} and σ_{yz}. The nine stress components completely define the state of stress of a body. They are described conveniently by the *stress matrix*,

$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix}. \tag{5.3}$$

Imagine that you are pushing an object with both hands. If each hand pushes equally strongly, the object will simply be displaced (the motion is called translation). However, if one hand pushes harder than the other, the unbalanced forces will cause the object to rotate as well as be translated. Applying these considerations to the stress matrix, if the forces on a body are balanced so as to give no rotation, the 3 × 3 matrix is symmetric (i.e., $\sigma_{xy} = \sigma_{yx}$; $\sigma_{yz} = \sigma_{zy}$; $\sigma_{zx} = \sigma_{xz}$) and contains only six independent elements.

5.1.3 The Strain Matrix

5.1.3.1 Longitudinal Strain

The strains produced in a body can also be expressed by a 3 × 3 matrix. Consider first the one-dimensional case shown in Fig. 5.3 of two points in a body located close together at the positions x and $(x + \Delta x)$. If the point x is displaced by an infinitesimally small amount u in the direction of the x-axis, the point $(x + \Delta x)$ will be displaced by $(u + \Delta u)$, where Δu is equal to $(\partial u/\partial x)\Delta x$ to first order. The *longitudinal strain* or *extension* in the x-direction is the fractional change in length of an element along the x-axis. The original separation of the two points was Δx; one point was displaced by u, the other by $(u + \Delta u)$, so the new separation of the points is $(\Delta x + \Delta u)$. The component of strain parallel to the x-axis resulting from a small displacement parallel to the x-axis is denoted ε_{xx}, and is given by

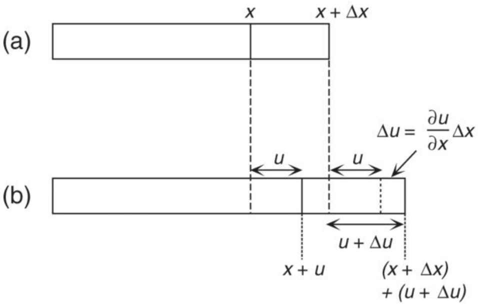


Fig. 5.3 Infinitesimal displacements u and $(u + \Delta u)$ of two points in a body that are located close together at the positions x and $(x + \Delta x)$, respectively.

$$\varepsilon_{xx} = \frac{\left(\Delta x + \frac{\partial u}{\partial x}\Delta x\right) - \Delta x}{\Delta x} = \frac{\partial u}{\partial x}. \tag{5.4}$$

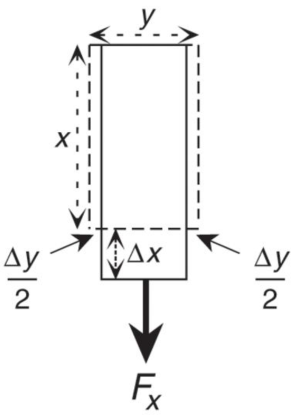
The description of longitudinal strain can be expanded to three dimensions. If a point (x, y, z) is displaced by an infinitesimal amount to $(x + u, y + v, z + w)$, two further longitudinal strains ε_{yy} and ε_{zz} are defined by

$$\varepsilon_{yy} = \frac{\partial v}{\partial y} \qquad \text{and} \qquad \varepsilon_{zz} = \frac{\partial w}{\partial z}. \tag{5.5}$$

In an elastic body the transverse strains ε_{yy} and ε_{zz} are not independent of the strain ε_{xx}. Consider the change of shape of the bar in Fig. 5.4. When it is stretched parallel to the *x*-axis, it becomes thinner parallel to the *y*-axis and parallel to the *z*-axis. The transverse longitudinal strains ε_{yy} and ε_{zz} are of opposite sign but proportional to the extension ε_{xx} and can be expressed as

$$\varepsilon_{yy} = -\nu\varepsilon_{xx} \qquad \text{and} \qquad \varepsilon_{zz} = -\nu\varepsilon_{xx}. \tag{5.6}$$

The constant of proportionality ν is called *Poisson's ratio*. The values of ν for all geophysically relevant materials lie between 0 (no lateral contraction) and a maximum value of 0.5 for an incompressible medium that does not experience any volume change upon deformation. In very hard, rigid rocks like granite, ν is about 0.45, while in soft, poorly consolidated sediments it is about 0.05. In the interior of the Earth, ν commonly has a value around 0.24–0.27. A body for which the value of ν equals 0.25 is called an ideal Poisson body. A few natural materials, such as cork, and some engineered materials, such as carbon nano-tubes, can have slightly negative Poisson ratios, meaning that they become thicker in response to stretching.



Poisson's ratio:

$$\nu = -\frac{\Delta y / y}{\Delta x / x} = -\frac{\varepsilon_{yy}}{\varepsilon_{xx}}$$

Fig. 5.4 Change of shape of a rectangular bar under extension. When stretched parallel to the *x*-axis, it becomes thinner parallel to the *y*-axis and *z*-axis.

5.1.3.2 Dilatation

The *dilatation* θ is defined as the fractional change in volume of an element in the limit when its surface area decreases to zero. Consider an undeformed volume element (as in the description of longitudinal strain) which has sides Δx, Δy, and Δz and undistorted volume $V = \Delta x\, \Delta y\, \Delta z$. As a result of the infinitesimal displacements Δu, Δv and Δw the edges increase to $\Delta x + \Delta u$, $\Delta y + \Delta v$, and $\Delta z + \Delta w$, respectively. The fractional change in volume is

$$\begin{aligned}\frac{\Delta V}{V} &= \frac{(\Delta x + \Delta u)(\Delta y + \Delta v)(\Delta z + \Delta w) - \Delta x \Delta y \Delta z}{\Delta x \Delta y \Delta z} \\ &\approx \frac{\Delta x \Delta y \Delta z + \Delta u \Delta y \Delta z + \Delta v \Delta z \Delta x + \Delta w \Delta x \Delta y - \Delta x \Delta y \Delta z}{\Delta x \Delta y \Delta z} \\ &= \frac{\Delta u}{\Delta x} + \frac{\Delta v}{\Delta y} + \frac{\Delta w}{\Delta z},\end{aligned} \tag{5.7}$$

where very small products like $\Delta u\,\Delta v$, $\Delta v\,\Delta w$, $\Delta w\,\Delta u$, and $\Delta u\,\Delta v\,\Delta w$ have been ignored. In the limit, as Δx, Δy, and Δz all approach zero we get the dilatation

$$\theta = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}$$

$$\theta = \varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}. \tag{5.8}$$

5.1.3.3 Shear Strain

During deformation, a body generally experiences not only longitudinal strains as described above. The shear components of stress (σ_{xy}, σ_{yz}, σ_{zx}) produce *shear strains*, which are manifest as changes in the angular relationships between parts of a body. This is most easily illustrated in two dimensions. Consider a rectangle ABCD with sides Δx and Δy and its distortion due to shear stresses acting in the x–y plane (Fig. 5.5). As in the earlier example of longitudinal strain, the point A is displaced parallel to the x-axis by an amount u (Fig. 5.5a). Because of the shear deformation, points between A and D experience larger x-displacements the further they are from A. The point D which is at a vertical distance Δy above A is displaced by the amount $(\partial u/\partial y)\,\Delta y$ in the direction of the x-axis. This causes a clockwise rotation of side AD through a small angle ϕ_1 given by

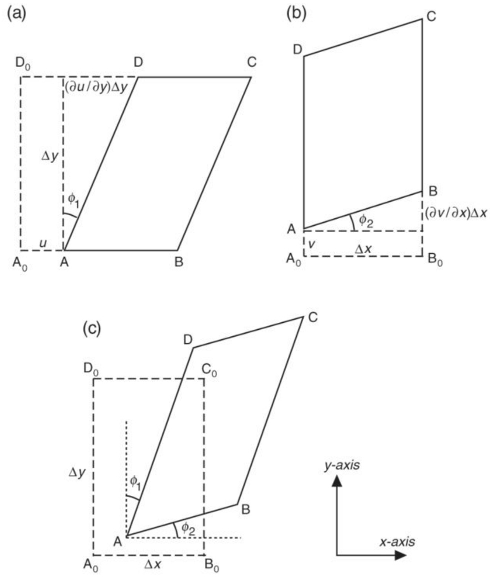

(a)
D_0
D
C
$(\partial u/\partial y)\Delta y$
Δy
ϕ_1
u
A_0
A
B
(b)
C
D
B
$(\partial v/\partial x)\Delta x$
ϕ_2
A
v
Δx
A_0
B_0
(c)
C
D
D_0
C_0
Δy
ϕ_1
B
ϕ_2
A
A_0
Δx
B_0
y-axis
x-axis

Fig. 5.5 (a) When a square is sheared parallel to the *x*-axis, side AD parallel to the *y*-axis rotates through a small angle ϕ_1; (b) when it is sheared parallel to the *y*-axis, side AB parallel to the *x*-axis rotates through a small angle ϕ_2. In general, shear causes both sides to rotate, giving a total angular deformation ($\phi_1 + \phi_2$). In each case the diagonal AC is extended.

$$\tan\phi_1 = \frac{(\partial u/\partial y)\Delta y}{\Delta y} = \frac{\partial u}{\partial y}. \tag{5.9}$$

Similarly, point A is displaced parallel to the *y*-axis by an amount *v* (Fig. 5.5b), while point B, which is at a horizontal distance Δx from A, is displaced by the amount $(\partial v/\partial x)\,\Delta x$ in the direction of the *y*-axis. As a result, side AB rotates counterclockwise through a small angle ϕ_2 given by

$$\tan\phi_2 = \frac{(\partial v/\partial x)\Delta x}{\Delta x} = \frac{\partial v}{\partial x}. \tag{5.10}$$

Elastic deformation involves infinitesimally small displacements and distortions, and for small angles we can write tan $\phi_1 = \phi_1$ and tan $\phi_2 = \phi_2$. The shear strain in the x–y plane (ε_{xy}) is defined as half the total angular distortion (Fig. 5.5c):

$$\varepsilon_{xy} = \frac{1}{2}\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right). \tag{5.11}$$

By transposing x and y, and the corresponding displacements u and v, the shear component ε_{yx} is obtained:

$$\varepsilon_{yx} = \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right). \tag{5.12}$$

This is identical to ε_{xy}. The total angular distortion in the x–y plane is $(\varepsilon_{xy} + \varepsilon_{yx}) = 2\varepsilon_{xy} = 2\varepsilon_{yx}$. Similarly, strain components $\varepsilon_{yz}\ (= \varepsilon_{zy})$ and $\varepsilon_{xz}\ (= \varepsilon_{zx})$ are defined for angular distortions in the y–z and z–x planes, respectively:

$$\begin{aligned}\varepsilon_{yz} = \varepsilon_{zy} &= \frac{1}{2}\left(\frac{\partial w}{\partial y} + \frac{\partial v}{\partial z}\right)\\ \varepsilon_{zx} = \varepsilon_{xz} &= \frac{1}{2}\left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}\right).\end{aligned} \tag{5.13}$$

The longitudinal and shear strains define the symmetric 3 × 3 *strain matrix*,

$$\begin{bmatrix} \varepsilon_{xx} & \varepsilon_{xy} & \varepsilon_{xz} \\ \varepsilon_{yx} & \varepsilon_{yy} & \varepsilon_{yz} \\ \varepsilon_{zx} & \varepsilon_{zy} & \varepsilon_{zz} \end{bmatrix}. \tag{5.14}$$

5.1.4 The Elastic Moduli

According to Hooke's law, when a body deforms elastically there is a linear relationship between stress and strain. The ratio of stress to strain defines an elastic modulus (or elastic parameter) of the body. Strain is itself a ratio of lengths and therefore dimensionless. Thus the elastic moduli must have the units of stress (N m^{-2}). Specifically defined elastic moduli, such as Young's modulus, the rigidity modulus, and the bulk modulus, describe the deformation of a material in response to a specific configuration of forces. As we will see, these moduli can generally be expressed in terms of each other.

Young's modulus is defined from the extensional deformations. Each longitudinal strain is proportional to the corresponding stress component; that is,

$$\sigma_{xx} = E\varepsilon_{xx} \qquad \sigma_{yy} = E\varepsilon_{yy} \qquad \sigma_{zz} = E\varepsilon_{zz}, \tag{5.15}$$

where the constant of proportionality, E, is the Young's modulus.

The rigidity modulus (or *shear modulus*), μ, is defined from the shear deformation. Like the longitudinal strains, the total shear strain in each plane is proportional to the corresponding shear stress component:

$$\sigma_{xy} = 2\mu\varepsilon_{xy} \qquad \sigma_{yz} = 2\mu\varepsilon_{yz} \qquad \sigma_{zx} = 2\mu\varepsilon_{zx}, \tag{5.16}$$

where μ is the proportionality constant The factor 2 arises as explained for Eqs. (5.11) and (5.12).

The *bulk modulus* (or *incompressibility*) is defined from the dilatation experienced by a body under hydrostatic pressure. Shear components of stress are zero for hydrostatic conditions ($\sigma_{xy} = \sigma_{yz} = \sigma_{zx} = 0$), and the normal stresses in all directions are equal to the negative hydrostatic pressure, p, ($\sigma_{xx} = \sigma_{yy} = \sigma_{zz} = -p$).

The bulk modulus, K, is the negative ratio of p to the dilatation; that is,

$$p = -K\theta.$$

(5.17)

The inverse of the bulk modulus (K^{-1}) is called the *compressibility*.

5.1.4.1 Bulk Modulus in Terms of Young's Modulus and Poisson's Ratio

Consider a rectangular volume element subjected to normal stresses σ_{xx}, σ_{yy}, and σ_{zz} on its end surfaces. Each longitudinal strain ε_{xx}, ε_{yy}, and ε_{zz} results from the combined effects of σ_{xx}, σ_{yy}, and σ_{zz} . For example, applying Hooke's law, the stress σ_{xx} produces an extension equal to σ_{xx}/E in the x-direction. The stress σ_{yy} causes an extension σ_{yy}/E in the y-direction, which results in an accompanying transverse strain $-\nu\ (\sigma_{yy}/E)$ in the x-direction, where ν is Poisson's ratio. Similarly, the stress component σ_{zz} makes a contribution $-\nu\ (\sigma_{zz}/E)$ to the total longitudinal strain ε_{xx} in the x-direction. Therefore,

$$\varepsilon_{xx} = \frac{\sigma_{xx}}{E} - \nu\frac{\sigma_{yy}}{E} - \nu\frac{\sigma_{zz}}{E}.$$

(5.18)

Similar equations describe the total longitudinal strains ε_{yy} and ε_{zz}. They can be rearranged as

$$\begin{aligned} E\varepsilon_{xx} &= \sigma_{xx} - \nu\sigma_{yy} - \nu\sigma_{zz} \\ E\varepsilon_{yy} &= \sigma_{yy} - \nu\sigma_{zz} - \nu\sigma_{xx} \\ E\varepsilon_{zz} &= \sigma_{zz} - \nu\sigma_{xxy} - \nu\sigma_{yy}. \end{aligned}$$

(5.19)

Adding these three equations together, we get

$$E(\varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}) = (1 - 2\nu)(\sigma_{xx} + \sigma_{yy} + \sigma_{zz}).$$

(5.20)

Consider now the effect of a constraining hydrostatic pressure, p, where $\sigma_{xx} = \sigma_{yy} = \sigma_{zz} = -p$. Using the definition of dilatation (θ) in Eq. (5.8) we get

$$E\theta = (1 - 2\nu)(-3p)$$

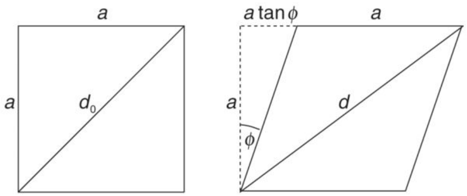
$$E = (1 - 2\nu)\left(-3\frac{p}{\theta}\right), \tag{5.21}$$

from which, using the definition of bulk modulus (K) in Eq. (5.17),

$$K = \frac{E}{3(1 - 2\nu)}. \tag{5.22}$$

5.1.4.2 Shear Modulus in Terms of Young's Modulus and Poisson's Ratio

The relationship between μ and E can be appreciated by considering the shear deformation of a rectangular prism that is infinitely long in one dimension and has a square cross-section in the plane of deformation. The shear causes shortening of one diagonal and extension of the other. Let the length of the side of the square be a (Fig. 5.6a) and that of its diagonal be d_0 ($= a\sqrt{2}$). The small shear through the angle ϕ displaces one corner by the amount ($a \tan \phi$) and stretches the diagonal to the new length d (Fig. 5.6b), which is given by Pythagoras' theorem:



(a) (b)

Fig. 5.6 (a) In the undeformed state d_0 is the length of the diagonal of a square with side length a. (b) When the square is deformed by shear through an angle ϕ, the diagonal is extended to the new length *d*.

$$\begin{aligned} d^2 &= a^2 + (a + a\tan\phi)^2 \\ &= a^2 + a^2(1 + 2\tan\phi + \tan^2\phi) \\ &= 2a^2\left(1 + \tan\phi + \frac{1}{2}\tan^2\phi\right) \approx d_0^2(1 + \phi), \end{aligned}$$

(5.23)

$$d \approx d_0\left(1 + \frac{1}{2}\phi\right),$$

(5.24)

where for an infinitesimally small strain tan ϕ = ϕ, and powers of ϕ higher than first order are negligibly small. The extension of the diagonal is

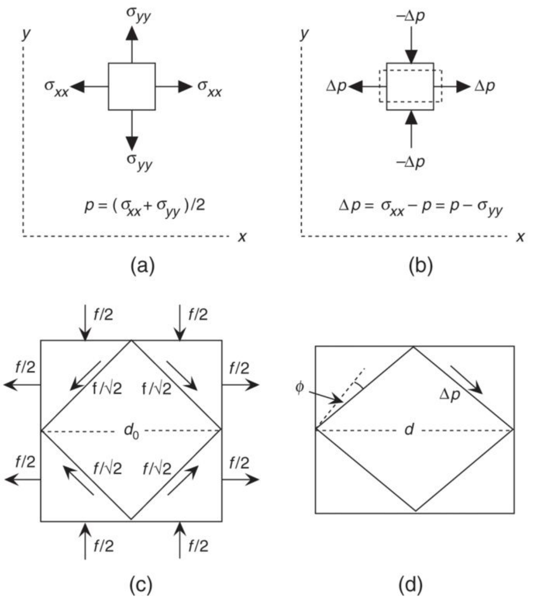
$$\frac{\Delta d}{d_0} = \frac{d - d_0}{d_0} = \frac{\phi}{2}.$$

(5.25)

This extension is related to the *normal* stresses σ_{xx} and σ_{yy} in the *x*–*y* plane of the cross-section (Fig. 5.7a), which are in general unequal. Let *p* represent their average value: $p = (\sigma_{xx} + \sigma_{yy})/2$. The change of shape of the square cross-section results from the differences Δ*p* between *p* and σ_{xx} and σ_{yy}, respectively (Fig. 5.7b). The outwards stress difference Δ*p* along the *x*-axis produces an *x*-extension equal to Δ*p*/*E*, while the inwards stress difference along the *y*-axis causes contraction along the *y*-axis and a corresponding contribution to the *x*-extension equal to ν (Δ*p*/*E*), where ν is Poisson's ratio as before. The total *x*-extension Δ*x*/*x* is therefore given by

$$\frac{\Delta x}{x} = \frac{\Delta p}{E}(1 + \nu).$$

(5.26)



σ_{yy}
y
σ_{xx}
σ_{xx}
σ_{yy}
$p = (\sigma_{xx} + \sigma_{yy})/2$
x
(a)
$-\Delta p$
y
Δp
Δp
$-\Delta p$
$\Delta p = \sigma_{xx} - p = p - \sigma_{yy}$
x
(b)
$f/2$
$f/2$
$f/2$
$f/2$
$f/\sqrt{2}$
$f/\sqrt{2}$
d_0
$f/2$
$f/2$
$f/\sqrt{2}$
$f/\sqrt{2}$
$f/2$
$f/2$
(c)
ϕ
Δp
d
(d)

Fig. 5.7 (a) Unequal *normal* stresses σ_{xx} and σ_{yy} in the x–y plane, and their average value p. (b) Stress differences Δp between p and σ_{xx} and σ_{yy}, respectively, cause elongation parallel to x and shortening parallel to y. (c) Forces $f = \Delta pA$ along the sides of the original square give shear forces $f/\sqrt{2}$ along the edges of the inner square, each of which has area $A\sqrt{2}$. (d) The shear stress on each side of the inner square has value Δp and causes extension of the diagonal of the inner square and shear deformation through an angle ϕ.

Let each edge of the square represent an arbitrary area A normal to the plane of the figure. The stress differences Δp produce forces $f = \Delta pA$ on the edges of the square, which resolve to shear forces $f/\sqrt{2}$ parallel to the sides of the inner square defined by joining the midpoints of the sides of the original square (Fig. 5.7c). Normal to the plane of the figure the surface area represented by each inner side is $A/\sqrt{2}$, and therefore the tangential (shear) stress acting on these sides simply equals Δp (Fig. 5.7d). The inner square shears through an angle ϕ, and so we can write

$$\Delta p = \mu\phi. \tag{5.27}$$

One diagonal becomes stretched in the *x*-direction while the other diagonal is shortened in the *y*-direction. The extension of the diagonal of a sheared square was shown above to be ϕ/2. Thus,

$$\frac{\Delta p}{E}(1+\nu) = \frac{\phi}{2} = \frac{\Delta p}{2\mu}. \tag{5.28}$$

Rearranging terms we get the relationship between *μ, E*, and ν:

$$\mu = \frac{E}{2(1+\nu)}. \tag{5.29}$$

5.1.4.3 The Lamé Constants

The first line of Eq. (5.19) can be rewritten as

$$E\varepsilon_{xx} = (1+\nu)\sigma_{xx} - \nu(\sigma_{xx} + \sigma_{yy} + \sigma_{zz}), \tag{5.30}$$

and from Eq. (5.20) we have

$$\begin{aligned}(\sigma_{xx} + \sigma_{yy} + \sigma_{zz}) &= \frac{E}{(1-2\nu)}(\varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}) \\ &= \frac{E}{(1-2\nu)}\theta,\end{aligned} \tag{5.31}$$

where *θ* is the dilatation, as defined in Eq. (5.8). After substituting Eq. (5.31) in Eq. (5.30) and rearranging we get

$$\sigma_{xx} = \frac{\nu E}{(1+\nu)(1-2\nu)}\theta + \frac{E}{(1+\nu)}\varepsilon_{xx}. \tag{5.32}$$

Writing V and substituting from Eq. (5.29), we can write Eq. (5.32) in the simpler form

$$\sigma_{xx} = \lambda\theta + 2\mu\varepsilon_{xx}, \tag{5.33}$$

with similar expressions for σ_{yy} and σ_{zz}.

The constants λ and μ are known as the *Lamé constants*. They are related to the elastic constants defined physically above. The modulus μ is equivalent to the rigidity modulus, while the bulk modulus K, Young's modulus E, and Poisson's ratio ν can each be expressed in terms of both λ and μ (Box 5.1).

Box 5.1 Elastic Parameters in Terms of the Lamé Constants

1. Bulk Modulus (*K*)

The bulk modulus describes volumetric shape changes of a material under the effects of the *normal* stresses σ_{xx}, σ_{yy}, and σ_{zz}. Writing Hooke's law for each normal stress gives

$$\sigma_{xx} = \lambda\theta + 2\mu\varepsilon_{xx}, \tag{1a}$$

$$\sigma_{yy} = \lambda\theta + 2\mu\varepsilon_{yy}, \tag{1b}$$

$$\sigma_{zz} = \lambda\theta + 2\mu\varepsilon_{zz}. \tag{1c}$$

Adding Equations (1a), (1b), and (1c) together gives

$$\sigma_{xx} + \sigma_{yy} + \sigma_{zz} = 3\lambda\theta + 2\mu(\varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}).$$

(2)

The dilatation θ is defined by Eq. (5.8) as

$$(\varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}) = \theta.$$

(3)

For hydrostatic conditions we can write $\sigma_{xx} = \sigma_{yy} = \sigma_{zz} = -p$ and substitute in Eq. (2), which can now be rearranged in the form

$$-3p = 3\lambda\theta + 2\mu\theta.$$

(4)

Using the definition of the bulk modulus as $K = -p/\theta$, we get the result

$$K = \lambda + \frac{2}{3}\mu.$$

(5)

2. Young's Modulus (*E*)

Young's modulus describes the longitudinal strains when a uniaxial normal stress is applied to a material. When only the longitudinal stress σ_{xx} is applied (i.e., $\sigma_{yy} = \sigma_{zz} = 0$), Hooke's law becomes

$$\sigma_{xx} = \lambda\theta + 2\mu\varepsilon_{xx},$$

(6a)

$$0 = \lambda\theta + 2\mu\varepsilon_{yy},$$

(6b)

$$0 = \lambda\theta + 2\mu\varepsilon_{zz}.$$

(6c)

Adding Equations (6a), (6b), and (6c) together gives

$$\sigma_{xx} = 3\lambda\theta + 2\mu(\varepsilon_{xx} + \varepsilon_{yy} + \varepsilon_{zz}) = 3\lambda\theta + 2\mu\theta,$$

(7)

$$\sigma_{xx} = (3\lambda + 2\mu)\theta,$$

(8)

$$\theta = \frac{\sigma_{xx}}{(3\lambda + 2\mu)}.$$

(9)

This expression is now substituted in Eq. (6a), which becomes

$$\sigma_{xx} = \lambda\frac{\sigma_{xx}}{(3\lambda + 2\mu)} + 2\mu\varepsilon_{xx}.$$

(10)

Gathering and rearranging terms gives the following succession:

$$\sigma_{xx}\left(1-\frac{\lambda}{3\lambda+2\mu}\right)=2\mu\varepsilon_{xx},$$

(11)

$$\sigma_{xx}\left(\frac{\lambda+\mu}{3\lambda+2\mu}\right)=\mu\varepsilon_{xx},$$

(12)

$$\sigma_{xx}=\mu\left(\frac{3\lambda+2\mu}{\lambda+\mu}\right)\varepsilon_{xx}.$$

(13)

The definition of Young's modulus is $E = \sigma_{xx}/\varepsilon_{xx}$ and so in terms of the Lamé constants,

$$E=\mu\left(\frac{3\lambda+2\mu}{\lambda+\mu}\right).$$

(14)

3. Poisson's Ratio (ν)

Poisson's ratio is defined as $\nu=-\varepsilon_{yy}/\varepsilon_{xx}=-\varepsilon_{zz}/\varepsilon_{xx}$. It relates the bulk modulus K and Young's modulus E as developed in Eq. (5.22):

$$K=\frac{E}{3(1-2\nu)}.$$

(15)

Substituting the derived expressions in Eq. (5) for K and Eq. (14) for E, we get

$$\frac{3\lambda + 2\mu}{3} = \frac{1}{3(1-2\nu)}\mu\left(\frac{3\lambda + 2\mu}{\lambda + \mu}\right).$$

(16)

This simplifies immediately to:

$$\frac{\lambda + \mu}{\mu} = \frac{1}{(1-2\nu)},$$

(17)

$$(1-2\nu) = \frac{\mu}{\lambda + \mu}.$$

(18)

Poisson’s ratio ν in terms of the Lamé constants is therefore given by

$$\nu = \frac{\lambda}{2(\lambda + \mu)}.$$

(19)

The values of **λ** and μ are almost equal in some materials, and it is often possible to assume $\boldsymbol{\lambda} = \mu$, from which it follows that $\nu = 0.25$. Such a material is called a *Poisson solid*.

5.1.4.4 Anisotropy

The elastic moduli are generally dependent on pressure, temperature, and the chemical composition of a material. Variations of temperature and pressure in the Earth ensure that the elastic moduli vary with depth. Additional variations are the result of differentiation, which was responsible for the formation of the Earth's internal structure.

In the foregoing discussion, it has been assumed that the relationships between stress and strain hold equally for all directions, a property called *isotropy*. This condition is not fulfilled in many minerals. For example, if a mineral has uniaxial symmetry in the arrangement of the atoms in its unit cell, the physical properties of the mineral parallel and perpendicular to the axis of symmetry are different. The mineral is *anisotropic*. The relations between components of stress and strain in an anisotropic substance are more complex than in the isotropic case examined in this chapter. The elastic moduli of an isotropic body are fully specified by the two moduli λ and μ or any other combination of two elastic moduli. However, as many as 21 moduli may be needed to describe anisotropic elastic behavior. In an anisotropic medium, the propagation speed of seismic waves, which depends on the elastic moduli, depends on the direction in which the wave is traveling.

Normally, a rock contains so many minerals that it may be assumed that they are oriented at random, and the rock may be treated as isotropic. This assumption may also be made for large regions of the Earth's interior. However, if anisotropic minerals are subjected to stress they develop a preferred alignment with the stress field, referred to as *lattice preferred orientation* (LPO). For example, platy minerals tend to align with their tabular shapes normal to the compression axis, or parallel to the direction of flow of a fluid. Preferential grain alignment results in seismic anisotropy. This has been observed in seismic studies of the upper mantle, especially at oceanic ridges, where anisotropic velocities have been attributed to the alignment of crystals by convection currents.

5.2 Viscous Flow

5.2.1 Solid-State Flow

While rocks behave approximately like elastic solids at near-surface conditions, the existence of plate tectonics suggests that parts of the Earth must have a different rheology when studied over geologic timescales. In fact, most of the mantle seems to flow even though temperatures at greater depth are not generally sufficient to melt large volumes of material; that is, to transform a solid crystalline material into a liquid.

Liquids, like water, flow because its constituent particles, such as H_2O molecules, are free to move. In contrast, the atoms of a crystalline solid are much less mobile. They are attached to their neighbors by chemical bonds that are difficult to break. Consequently, the mechanisms that enable *solid-state flow* must be very different from the deformation mechanism in a liquid.

To understand solid-state flow, we will first consider the physics of flow in liquids that we are more familiar with (Section 5.2.2). We will then see that solids do not flow by means of the movement of atoms but – in an analogous manner – by the propagation of defects in the crystal lattice. This phenomenon is known as *creep* (Section 5.2.3).

5.2.2 Viscous Flow in Liquids

Consider the case when a liquid or gas flows in thin layers parallel to a flat surface (Fig. 5.8). The *laminar flow* exists as long as the speed stays below a critical value, above which the flow becomes turbulent. Turbulent flow does not interest us here, because the rates of flow in solid Earth materials are very low.

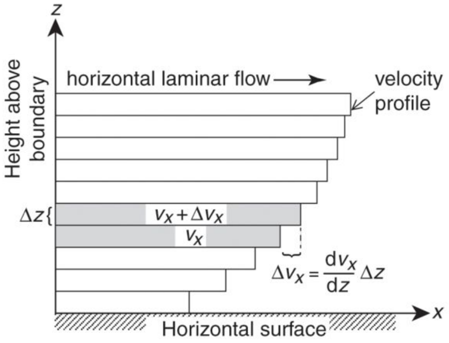


Fig. 5.8 Schematic representation of laminar flow of a fluid in infinitesimally thin layers parallel to a horizontal surface.

Suppose that the velocity of laminar flow along the horizontal x-direction increases with vertical height z above the reference surface. The molecules of the fluid may be regarded as having two components of velocity. One component is the velocity of flow in the x-direction, but in addition there is a random component with a variable velocity whose root mean square value determines the temperature (Section 9.2). Because of the random component, one-sixth of the molecules in a unit volume are moving upward and one-sixth downward on average at any time. This causes a transfer of molecules between adjacent layers in the laminar flow. The number of transfers per second depends on the size of the random velocity component, i.e., on temperature. The influx of molecules from the slower-moving layer reduces the momentum of the faster-moving layer. In turn, molecules transferred downward from the faster-velocity layer increase the momentum of the slower-velocity layer. This means that the two layers do not move freely past each other. They exert a shear force – or drag – on each other and the fluid is said to be *viscous*.

The magnitude of the shear force F_{xz} depends on how much momentum is transferred from one layer to the next. If all the molecules in a fluid have the same mass, the momentum transfer is determined by the change in the velocity v_x between the layers; this depends on the vertical gradient of the flow velocity (dv_x/dz). The momentum exchange depends also on the number of molecules that cross the boundary between adjacent layers and so is proportional to the surface area, A. We can bring these observations together, as did Newton in the seventeenth century, and derive the following proportionality relationship for F_{xz}:

$$F_{xz} \propto A \frac{dv_x}{dz}. \tag{5.34}$$

If we divide both sides by the area A, the left side becomes the shear stress σ_{xz}. Introducing a proportionality constant η, we get the equation

$$\sigma_{xz} = \eta \frac{dv_x}{dz}. \tag{5.35}$$

This equation is Newton's law of viscous flow, and η is the *viscosity*. If η does not depend on the flow velocity, the fluid is called a *Newtonian fluid*. The value of η depends on the transfer rate of molecules between layers, and so on temperature. Substituting the units of stress (pascal) and velocity gradient ($(\mathrm{m\ s^{-1}})/\mathrm{m} = \mathrm{s^{-1}}$), we find that the unit of η is a pascal-second (Pa s). A shear stress applied to a fluid with low viscosity causes a large velocity gradient; the fluid flows easily. This is the case in a gas (η in air is of the order 2×10^{-5} Pa s) or in a liquid (η in water is 1.005×10^{-3} Pa s at 20°C). The same shear stress applied to a very viscous fluid (with a large value of η) produces only a small velocity gradient transverse to the flow direction. The layers in the laminar flow are reluctant to move past each other. The viscous fluid is “sticky” and it resists flow. For example, η in a viscous liquid like engine oil is around 0.1–10 Pa s, three or four orders of magnitude higher than in water.

5.2.3 The Mechanisms of Flow in Solids

5.2.3.1 Crystals and Crystal Defects

The atoms in a perfect metal or crystal are arranged regularly to form a lattice with a simple symmetry. In some common arrangements the atoms are located at the corners of a cube or a hexagonal prism, defining a *unit cell* of the crystal. The lattice is formed by stacking unit cells (Fig. 5.9a). Theoretical calculations show that the shear stress needed to overcome atomic bonding in a perfect crystal and to make it flow viscously would be around 10–50 GPa for most silicates and metals in the Earth. This is much higher than the required shear stresses of less than 1 GPa that are found experimentally for real – that is, imperfect – Earth materials.

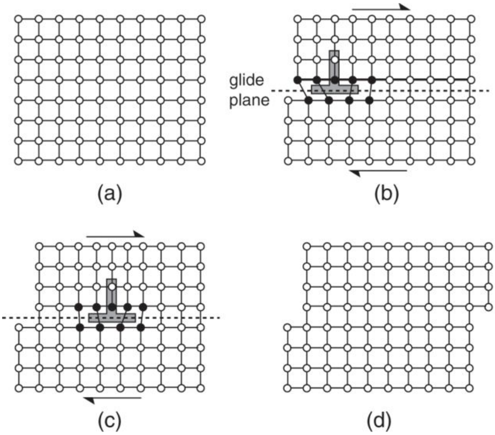


Fig. 5.9 Permanent shear deformation caused by the motion of an edge dislocation through a crystal: (a) undeformed crystal lattice, (b) arrangement of atoms around an edge dislocation in the lattice, (c) passage of the dislocation across the crystal, and (d) resulting sheared lattice. The head of the inverted T lies in the glide plane, its stalk marks the extra plane of atoms.

While liquids flow because of the mobility of their constituent molecules, the polycrystalline rocks that compose the Earth's crust and mantle flow of crystalline imperfections are known as *defects*. The simplest of all defects are missing atoms called *vacancies*. Similarly, *interstitial atoms* may occupy sites that are not occupied in an idealized, perfect crystal. Point defects such as vacancies and interstitial atoms may be distributed randomly throughout the crystal lattice, but they can also form long chains called *dislocations*.

There are several types of dislocations, the simplest being an edge dislocation. It is formed when an extra plane of atoms is introduced in the lattice (Fig. 5.9). The edge dislocation terminates at a plane perpendicular to it called the *glide plane*. Another common type of dislocation is the screw dislocation. It also is made up of atoms that are displaced from their regular positions, in this case forming a spiral about an axis. *Grain boundaries* in a rock composed of many individual crystal grains can also be considered a type of crystal defect or dislocation.

Generally, the atomic bonding around crystal defects is strongly disturbed, making atoms much more mobile than in a perfect single crystal. It is this defect-related mobility that enables viscous solid-state flow in the Earth.

5.2.3.2 Creep Mechanisms in the Earth

Crystal defects can be mobilized by shear stress. The migration of defects under stress is referred to as *creep*. It is the physical process behind the macroscopic phenomenon of solid-state flow.

While the palette of possible creep mechanisms is quite extensive, only two of them play a major role for Earth materials: *dislocation creep* and *diffusion creep*. Each mechanism is a thermally activated process. This means that the strain rate depends on the temperature T according to an exponential function with the form $e^{-E_a/kT}$. Here, k is Boltzmann's constant ($k = 1.38064852 \times 10^{-23}$ J K^{-1}), while E_a is the energy needed to activate the type of flow; it is called the *activation energy*. At low temperatures, where $T \ll E_a/k$, the strain rate is very low and creep is insignificant. Because of the exponential function, the strain rate increases rapidly with increasing temperature above T

$= E_a/k$. The type of flow at a particular depth depends on the local temperature and its relationship to the melting point temperature T_{mp}. Above T_{mp} the interatomic bonds in the solid break down and it flows as a true liquid.

Dislocation creep results from the application of external stress that mobilizes a dislocation, causing it to glide through the undisturbed part of the crystal (Fig. 5.9b–d). If the dislocation glide is not blocked by an obstacle, the dislocation migrates out of the crystal, leaving a shear deformation. The deformation rate that results from dislocation motion depends on the dislocation density, that is, the number and length of dislocations per volume. Since the application of stress tends to increase dislocation density by forming new dislocations, the stress–strain relationship for dislocation creep is a nonlinear power law, with strain rate $d\varepsilon/dt$ proportional to the nth power of stress σ:

$$\frac{d\varepsilon}{dt} = A\left(\frac{\sigma}{\mu}\right)^n e^{-E_a/kT}, \tag{5.36}$$

where μ is the rigidity modulus and A is a constant with the dimensions of strain rate; typically $n \geq 3$. This relationship means that the strain rate increases much more rapidly than the stress. Dislocation creep is the most important mechanism of flow at temperatures between $0.55\,T_{mp}$ and $0.85\,T_{mp}$. The temperature throughout most of the mantle exceeds half the melting point, and so dislocation creep is probably the flow mechanism that permits convection throughout most of the mantle. It is also likely to be the main form of deformation in the lower lithosphere, where the relationship of temperature to melting point is also suitable. An important aspect of dislocation creep is the reorganization of the grains, including the formation of new grain boundaries. Dislocation creep may reorient grains. This microscopic phenomenon, LPO, can result in observable seismic anisotropy when it occurs over large enough volumes.

Diffusion creep is the result of atoms moving from their stable positions to neighboring positions due to thermal vibrations. The driving forces for this atomic *diffusion* include stress gradients induced by an applied stress within a heterogeneous, polycrystalline material. Diffusion through the bulk of a crystal grain is referred to as *Nabarro–Herring creep*. In contrast, diffusion along grain boundaries is known as *Coble* creep.

In a real material, diffusion occurs preferentially along pre-existing grain boundaries and dislocations, suggesting that grain size L plays an important role in diffusion creep. Indeed, strain rate $d\varepsilon/dt$ is inversely proportional to L^m, with m ranging between 1 and 3, and linearly proportional to stress σ,

$$\frac{d\varepsilon}{dt} = A\frac{\sigma}{L^m}e^{-E_a/kT}, \tag{5.37}$$

where A is a proportionality constant. Equation (5.37) indicates that diffusion creep is most efficient when the grain size is small. For large grain sizes, dislocation creep tends to be dominant. Furthermore, the linear dependence of strain rate on stress shows that solid-state deformation by diffusion creep is indeed macroscopically identical to viscous flow of a Newtonian fluid, discussed in Section 5.2.2.

Experimental results suggest that diffusion creep may be the dominant creep mechanism in the deep parts of the Earth's upper mantle. In contrast to dislocation creep, diffusion creep does not lead to LPO of crystals, and therefore also not to seismic anisotropy.

5.3 Deviations from Perfect Elasticity and Viscous Flow

Perfect elastic deformation and viscous flow are idealizations of material deformation under stress. They are convenient for the description and understanding of numerous phenomena, ranging from the propagation of seismic waves to convection in the Earth's mantle and outer core. Often, however, heterogeneous Earth materials exhibit behavior that deviates from the idealization. For example, the deformation of rocks under small stress may be delayed in time or not be recoverable. When a large stress is applied, a rock may deform plastically, or it may fracture. These and other modes of non-ideal deformation will be discussed in the following paragraphs.

5.3.1 Anelasticity: Delayed Deformation

When a small stress σ is applied to the block in Fig. 5.1, its length will not extend instantaneously from h to $h + \Delta h$ because various processes require some time; atoms in the crystal move to new equilibrium positions, existing dislocations move, and new ones may form. The final elongation Δh is still proportional to σ, and the block fully recovers its original shape when the stress is removed. This kind of linear, recoverable, but non-instantaneous deformation is called *anelasticity*.

In 1890 Lord Kelvin modeled anelastic deformation by combining the characteristics of a perfectly elastic solid and a viscous liquid. An applied stress causes both elastic and viscous effects. If strain is ε, the corresponding elastic part of the stress is $E\varepsilon$, where E is Young's modulus (Section 5.1.4). Similarly, if the rate of change of strain with time is $d\varepsilon/dt$, the viscous part of the stress is $\eta\, d\varepsilon/dt$, where η is viscosity. The applied stress σ is the sum of the two parts and can be written

$$\sigma = E\varepsilon + \eta \frac{d\varepsilon}{dt}. \tag{5.38}$$

To solve this equation we first divide throughout by E, then define the *retardation time* or *Maxwell time* $\tau = \eta/E$, which is a measure of how long it takes for viscous strains to exceed elastic strains. Substituting and rearranging the equation, we get

$$\varepsilon + \tau \frac{d\varepsilon}{dt} = \frac{\sigma}{E}, \tag{5.39}$$

$$\frac{\varepsilon}{\tau}+\frac{d\varepsilon}{dt}=\frac{\varepsilon_m}{\tau}, \tag{5.40}$$

where $\varepsilon_m = \sigma/E$. Multiplying throughout by the integrating factor $e^{t/\tau}$ gives

$$\frac{\varepsilon}{\tau}e^{t/\tau}+\frac{d\varepsilon}{dt}e^{t/\tau}=\frac{\varepsilon_m}{\tau}e^{t/\tau}, \tag{5.41}$$

$$\frac{d}{dt}\left(\varepsilon \mathrm{e}^{t/\tau}\right)=\frac{\varepsilon_m}{\tau}\mathrm{e}^{t/\tau}. \tag{5.42}$$

Integrating both sides of this equation with respect to *t*, assuming a constant σ and therefore constant ε_m, gives

$$\varepsilon \mathrm{e}^{t/\tau}=\varepsilon_m \mathrm{e}^{t/\tau}+C, \tag{5.43}$$

where *C* is a constant of integration determined by the initial condition. If the initial strain is zero (i.e., at $t = 0$, $\varepsilon = 0$), substituting in Eq. (5.43) gives $C = \varepsilon_m$. The solution for the strain at time *t* is therefore

$$\varepsilon=\varepsilon_m\left(1-\mathrm{e}^{t/\tau}\right). \tag{5.44}$$

Thus, in contrast to elastic (instantaneous) behavior, the strain rises exponentially to a limiting value given by $\varepsilon_m = \sigma/E$.

5.3.2 Viscoelasticity: Delayed and Non-Recoverable Deformation

In addition to being anelastic, deformation can often not be recovered fully because crystal imperfections such as point defects and dislocations moved non-reversibly. Thus, when stress is removed, a permanent strain remains. Deformation that depends linearly on stress without being either recoverable or instantaneous is called *viscoelastic*.

Viscoelasticity combines perfectly elastic, anelastic, and viscous behavior. The relative importance of these components depends on the time during which stress is applied (Fig. 5.10a). At first, the rock at once strains elastically. This is followed by a stage in which the strain initially increases rapidly (i.e., the strain rate is high) and then levels off. If the stress is removed within this regime, the deformation drops rapidly to zero. This stage, characterized by anelastic behavior, is known as *primary creep* or *delayed elastic creep*. Beyond the primary stage, creep progresses at a slower and nearly constant rate by solid-state flow, as described in Section 5.2. This stage is called *secondary* or *steady-state creep*. When stress is removed, a permanent strain is left after the elastic and anelastic recoveries. Provided that the applied stress is sufficiently high, the strain rate increases ever more rapidly in a stage called *tertiary creep* or *plastic flow*, which may eventually lead to failure (see Sections 5.3.3 and 5.3.4).

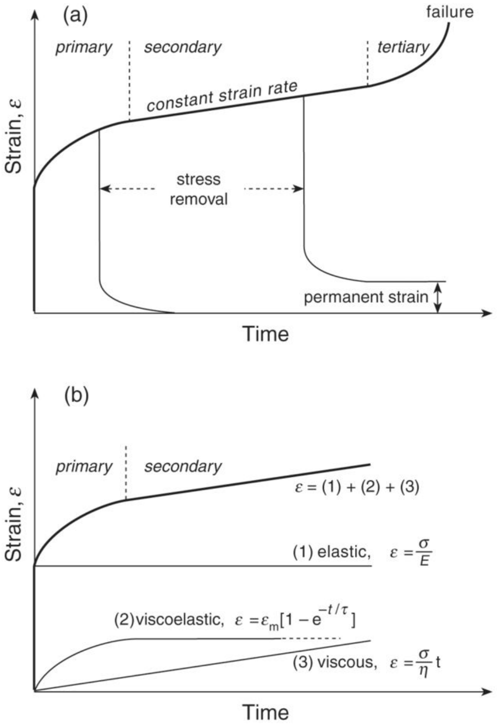


Fig. 5.10 (a) Schematic strain–time curve for a material deforming under constant stress, and (b) model creep curve that combines elastic, viscoelastic, and viscous elements (Eq. 5.45).

The primary and secondary stages of the creep curve in the presence of a constant stress can be modeled by combining elastic, anelastic, and viscous elements (Fig. 5.10b). The elastic component is given by $\varepsilon_{elastic} = \sigma/E$, and the anelastic component is $\varepsilon_{anelastic} = \varepsilon_m(1 - e^{-\tau/t})$, as in Eq. (5.44). The viscous component of strain rises linearly with time, corresponding to a constant strain rate. Combining terms gives the expression

$$\varepsilon = \frac{\sigma}{E} + \varepsilon_m\left(1 - e^{-t/\tau}\right) + \frac{\sigma}{\eta}t. \tag{5.45}$$

This simple model explains the main features of experimentally observed creep curves. It is in fact very difficult to ensure that laboratory observations are representative of creep in the Earth.

Conditions of pressure and temperature in the crust and upper mantle can be achieved or approximated. The major problems arise from the differences in timescales and creep rates. Even creep experiments conducted over months or years are far shorter than the lengths of time in a geological process. The creep rates in nature (e.g., around 10^{-14} s^{-1}) are many orders of magnitude slower than the slowest strain rate used in laboratory experiments (around 10^{-8} s^{-1}). Nevertheless, the experiments have provided a better understanding of the rheology of the Earth's interior and the physical mechanisms active at different depths.

The component of viscoelastic deformation that dominates not only depends on time but also on a large number of other factors. These include pressure and temperature, chemical composition, and the presence of cracks that may be filled with fluids.

In viscoelastic deformation part of the elastic energy is used to produce non-reversible reorganizations of the crystal lattice, including the movement of dislocations or the migration of point defects. As a consequence, the energy carried by seismic waves gradually attenuates as they propagate through the Earth. The attenuation of seismic waves ensures that the Earth does not vibrate eternally after an earthquake.

5.3.3 Plastic Flow

We have, so far, considered the deformation under small stress where a material may deform elastically (Hooke's law) or anelastically over short timescales. Beyond a certain value of the stress, called the *linearity limit* (Fig. 5.11), Hooke's law no longer holds. Although the material still returns to its original shape when stress is removed, the stress–strain relationship is nonlinear. If stress is further increased beyond a certain point, known as the *elastic limit* or *yield stress*, the material will not recover its original shape when stress is removed; a permanent strain has been produced. In this range a small increase in applied stress causes a disproportionately large increase in strain. The deformation is said to be *plastic*. The yield stress depends on numerous factors, including the composition of the material, pressure, and temperature. Near the Earth's surface, the yield stress of rocks is 10–100 MPa, as orders of magnitude.

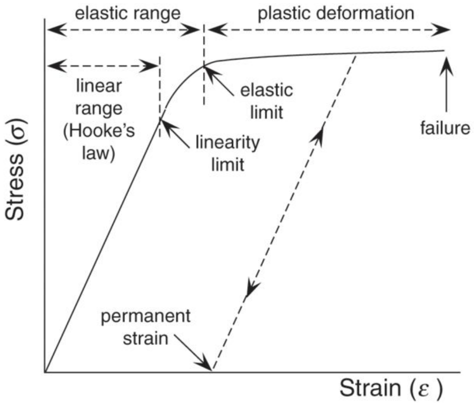


Fig. 5.11 (a) The stress–strain relation for a solid material is linear (Hooke's law) until the proportionality limit, and the material deforms elastically until it reaches the elastic limit. At sufficiently high stress, plastic deformation produces further strain until failure occurs.

Plastic deformation appears similar to solid-state viscous flow. The main difference is that viscous flow can result from a small stress, provided that it is applied over a long time at sufficiently high temperature. Plastic deformation, in contrast, requires stress to exceed the yield stress, which may happen over short times and at low temperatures. At the scale of a crystal, plastic flow may occur by the movement of dislocations. Since the stress is high, dislocations can move even at low temperatures where thermally activated dislocation creep (Section 5.2.3.2) is inefficient. As a consequence, the stress–strain relationship for plastic flow deviates from the dislocation creep power law (Eq. 5.36). This phenomenon is often referred to as *power-law breakdown*.

In *perfectly plastic* behavior the stress–strain curve has zero slope. However, the stress–strain curves of realistic plastically deformed materials usually have a small positive slope. This means that the stress must be further increased above the yield stress for plastic deformation to advance. This effect is called *strain-hardening*.

5.3.4 Fracture and the Brittle–Ductile Transition

Plastic flow at high stress and low temperature is a form a material *failure*. It is in contrast to failure by the formation of localized fractures, which is known as *brittle deformation. Brittle deformation* consists of rupture without other distortion. It is an abrupt process that causes faulting in rocks and earthquakes, accompanied by the release of elastic energy in the form of seismic waves. Brittle fracture occurs at much lower stresses than the intrinsic strength of a perfect crystal lattice. This is attributed to the presence of cracks, which modify the local internal stress field in the crystal. Fracture may occur under either extension or shear. Brittle deformation is the main mechanism in tectonic processes that involve the uppermost 5–10 km of the lithosphere.

Brittle deformation involves the complete breaking of atomic bonds and the formation of cracks. These processes generally involve significant volume expansion. As a consequence, pressure (depth) plays an important role. Above a certain pressure, cracks cannot open, and the brittle form of deformation transitions to *ductile deformation*, that is, deformation by solid-state viscous or plastic flow. The behavior of rocks and minerals in the deep interior of the Earth is characterized mainly by ductile deformation.

The transition from predominantly brittle to predominantly ductile deformation occurs differently in oceanic and continental lithosphere (Fig. 5.12). The depth of the transition depends on several parameters, including the composition of the rocks, the local geothermal gradient, crustal thickness, and strain rate. Consequently it is sensitive to the vertically layered structure of the lithosphere. Oceanic lithosphere has a thin crust and shows a nearly linear increase in strength (yield stress) with depth that results from the nearly linear increase in pressure. Strength reaches a maximum in the upper mantle at about 30–40 km depth, where the effect of increasing temperature starts to be dominant. At greater depths temperature is high enough for thermally activated diffusion and dislocation creep. Therefore, the lithosphere gradually becomes more ductile, eventually grading into the low-rigidity asthenosphere below about 100 km depth.

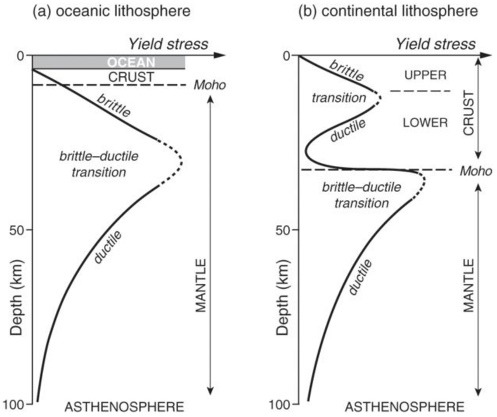

(a) oceanic lithosphere
Yield stress
OCEAN
CRUST
Moho
brittle
brittle–ductile transition
ductile
MANTLE
ASTHENOSPHERE
Depth (km)
0
50
100
(b) continental lithosphere
Yield stress
brittle
UPPER
transition
ductile
LOWER
CRUST
Moho
brittle–ductile transition
ductile
MANTLE
ASTHENOSPHERE
Depth (km)
0
50
100

Fig. 5.12 Schematic vertical profiles of yield stress in (a) oceanic lithosphere and (b) continental lithosphere with the estimated depths of brittle–ductile transitions. Alluding to its shape, the profile for continental lithosphere is colloquially referred to as a *Christmas tree diagram*.

Source: P. Molnar, Continental tectonics in the aftermath of plate tectonics, *Nature* 335, 131–137, 1988.

The continental crust is much thicker than the oceanic crust and has a more complex layering. The upper crust is brittle, deforming mostly by frictional sliding on faults. Initially, strength increases with increasing depth (pressure) until minerals are weakened sufficiently by high temperature. This first maximum in strength approximately marks the boundary between the brittle upper and lower crust, which becomes increasingly ductile toward the Moho at about 30–35 km depth, on average. In the upper mantle the strength increases again due to a change in mineral composition. This leads to a second brittle–ductile transition at an average depth of 40–50 km.

The difference in rheological layering of continental and oceanic lithosphere is important in collisions between plates. The crustal part of the continental lithosphere may detach from the mantle. The folding, underthrusting and stacking of crustal layers produce folded mountain ranges in the suture zone and thickening of the continental crust. For example, great crustal thicknesses exceeding 70 km under the Himalayas are attributed to underthrusting of crust from the Indian plate beneath the crust of the Eurasian plate.

5.4 Lithosphere Rigidity

Lithospheric plates are thin compared to their horizontal extents. However, they evidently react rigidly to the forces that propel them. The lithosphere does not easily buckle under horizontal stress. A simple analogy may be made to a thin sheet of paper resting on a flat pillow. If pushed on one edge, the page simply slides across the pillow without crumpling. Only if the leading edge encounters an obstacle does the page bend upward some distance in front of the hindrance, while the leading edge tries to burrow under it. This is what happens when an oceanic lithospheric plate collides with another plate. A small forebulge develops on the oceanic plate and the leading edge bends downward into the mantle, forming a subduction zone.

The ability to bend is a measure of the rigidity of the plate. This is also manifest in its reaction to a local vertical load. If, in our analogy, a small weight is placed in the middle of the page, it is pressed down into the soft pillow. A large area around the weight takes part in this process, which may be compared with the Vening Meinesz type of regional isostatic compensation (Section 4.3.2.3). The weight of a seamount or chain of islands has a similar effect on the oceanic lithosphere. By studying the flexure due to local vertical loads, information is obtained about a *static property* of the lithosphere, namely its resistance to bending.

In our analogy the locally loaded paper sheet would not bend if it lay on a hard, flat table. It is only able to flex if it rests on a soft, yielding surface. After the weight is removed, the page is restored to its original flat shape. The restoration after unloading is a measure of the properties of the pillow as well as the page. A natural example is the rebound of regions (such as the Canadian shield or Fennoscandia; see Section 4.3) that have been depressed by now-vanished ice-sheets. The analysis of the rates of glacial rebound provides information about a *dynamic property* of the mantle beneath the lithosphere. The depression of the surface forces mantle material to flow away laterally to make way for it; when the load is removed, the return mantle flow presses the concavity back upward. The ease with which the mantle material flows is described by its dynamic viscosity.

The resistance to bending of a thin, elastic plate overlying a weak fluid is expressed by an elastic parameter called the *flexural rigidity* and denoted *D*. For a plate of thickness *h*,

$$D = \frac{E}{12(1-\nu^2)} h^3, \tag{5.46}$$

where *E* is Young's modulus and ν is Poisson's ratio (see Section 5.1.4 for the definition of these elastic constants). The dimensions of *E* are N m^{-2} and ν is dimensionless; hence, the dimensions of *D* are those of a bending moment (N m). *D* is a fundamental parameter of the elastic plate, which describes how easily it can be bent. A large value of *D* corresponds to a stiff plate, which may result from a large Young's modulus or a large thickness.

Here we consider two situations of particular interest for the rigidity of the oceanic lithosphere. The first is the bending of the lithosphere by a topographic feature such as an oceanic island or seamount; only the vertical load on the elastic plate is important. The second is the bending of the lithosphere at a subduction zone. In this case vertical and horizontal forces are located along the edge of the plate, and the plate experiences a bending moment which deflects it downward.

5.4.1 Lithospheric Flexure Caused by Oceanic Islands

The theory for elastic bending of the lithosphere is derived from the bending of thin elastic plates and beams. This involves a fourth-order differential equation, whose derivation is beyond the scope of this book. However, it is instructive to consider the forces involved in setting up the equation, and to examine its solution in a simplified context.

Consider the bending of a thin isotropic elastic plate of thickness h carrying a surface load $L(x, y)$ and supported by a substratum of density ρ_m (Fig. 5.13a). Let the deflection of the plate at a position (x, y) relative to the center of the load be $w(x, y)$. Two forces counteract the downward force of the load. The first, arising from Archimedes' principle, is a buoyant force equal to $(\rho_m - \rho_i)\, gw$, where ρ_i is the density of the material that fills in the depression caused by the deflection of the plate. The second force arises from the elasticity of the beam. Elasticity theory shows that this produces a restoring force proportional to a fourth-order differential of the deflection w. Balancing the elastic and buoyancy forces against the deforming load leads to the equation

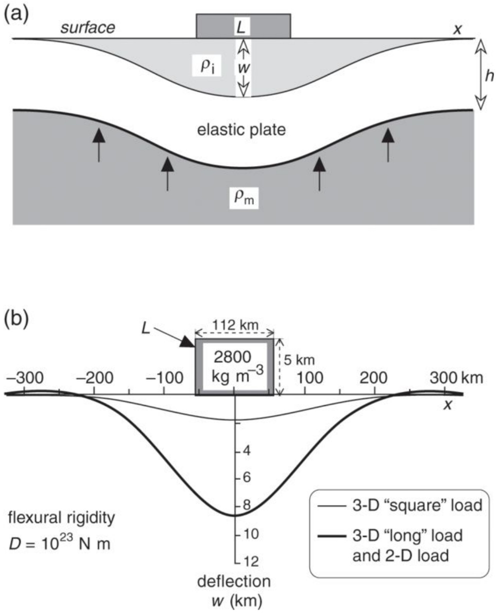

(a)
surface
L
x
ρi
w
h
elastic plate
ρm
(b)
L
112 km
2800 kg m−3
5 km
−300
−200
−100
100
200
300 km
x
4
6
8
10
12
deflection w (km)
flexural rigidity D = 10²³ N m
3-D "square" load
3-D "long" load and 2-D load

Fig. 5.13 (a) Geometry for the elastic bending of a thin plate of thickness *h* supported by a denser substratum. The surface load *L* causes a downward bending *w*. (b) Comparison of 2-D and 3-D elastic plate models. The load is taken to be a topographic feature of density 2800 kg/m³, height 5 km, and cross-sectional width 112 km.

Source: A.B. Watts, J.R. Cochran and G. Selzer, Gravity anomalies and flexure of the lithosphere: a three-dimensional study of the Great Meteor seamount, Northeast Atlantic, *J. Geophys. Res.* 80, 1391–1398, 1975.

$$D\left\{\frac{\partial^4 w}{\partial x^4}+2\frac{\partial^4 w}{\partial x^2 \partial y^2}+\frac{\partial^4 w}{\partial y^4}\right\}+(\rho_{\mathrm{m}}-\rho_{\mathrm{i}})gw=L(x,y). \tag{5.47}$$

For a linear topographic feature, such as a mountain range or chain of seamounts, the bending geometry is the same in any cross-section normal to its length and the problem reduces to the two-dimensional elastic bending of a thin beam. If the load is a linear feature in the *y*-direction, the variation of *w* with *y* disappears and the differential equation becomes:

$$D\frac{\partial^4 w}{\partial x^4}+(\rho_{\mathrm{m}}-\rho_{\mathrm{i}})gw=L.$$
(5.48)

An important example is a linear load *L* concentrated along the *y*-axis at $x = 0$. The solution is a damped sinusoidal function:

$$w=w_o e^{-x/\alpha}\left(\cos\frac{x}{\alpha}+\sin\frac{x}{\alpha}\right),$$
(5.49)

where w_0 is the amplitude of the maximum deflection underneath the load (at $x = 0$). The parameter α is called the *flexural parameter*; it is related to the flexural rigidity *D* of the plate by

$$\alpha^4=\frac{4D}{(\rho_{\mathrm{m}}-\rho_{\mathrm{i}})g}.$$
(5.50)

The elasticity of the plate (or beam) distributes the load of the surface feature over a large lateral distance. The fluid beneath the load is pushed aside by the penetration of the plate. The buoyancy of the displaced fluid forces it upward, causing uplift of the surface adjacent to the central depression. Equation (5.49) shows that this effect is repeated with increasing distance from the load, the wavelength of the fluctuation is $\lambda = 2\pi\alpha$. The amplitude of the disturbance diminishes rapidly because of the exponential attenuation factor. Usually it is only necessary to consider the central depression and the first uplifted region. The wavelength λ is equal to the distance across the central depression. Substituting in Eq. (5.50) gives *D*, which is then used with the parameters *E* and ν in Eq. (5.46) to obtain *h*, the thickness of the elastic plate. The computed values of *h* depend on the rheological details of the crust and upper mantle, and therefore usually do not agree with the thickness of the crust. For example, h exceeds the crustal thickness when the elastic plate includes part of the upper mantle. The value of h is equated with the thickness of the *elastic lithosphere*.

The difference between the deflection caused by a two-dimensional (2-D) load (i.e., a linear feature) and that due to a three-dimensional (3-D) load (i.e., a feature that has limited extent in the x- and y-directions) is illustrated in Fig. 5.13b. If the length of the 3-D load normal to the cross-section is more than about 10 times its width, the deflection is the same as for a 2-D load. A load with a square base (i.e., extending the same distance along both x- and y-axes) causes a central depression that is less than one-quarter the effect of the linear load.

The validity of the lithospheric flexural model of isostasy can be tested by comparing the computed gravity effect of the model with the observed free-air gravity anomaly Δg_F. The isostatic compensation of the Great Meteor seamount in the North Atlantic provides a suitable test for a 3-D model (Fig. 5.14). The shape of the free-air gravity anomaly obtained from detailed marine gravity surveys was found to be fitted best by a model in which the effective flexural rigidity of the deformed plate was assumed to be 6×10^{22} N m.

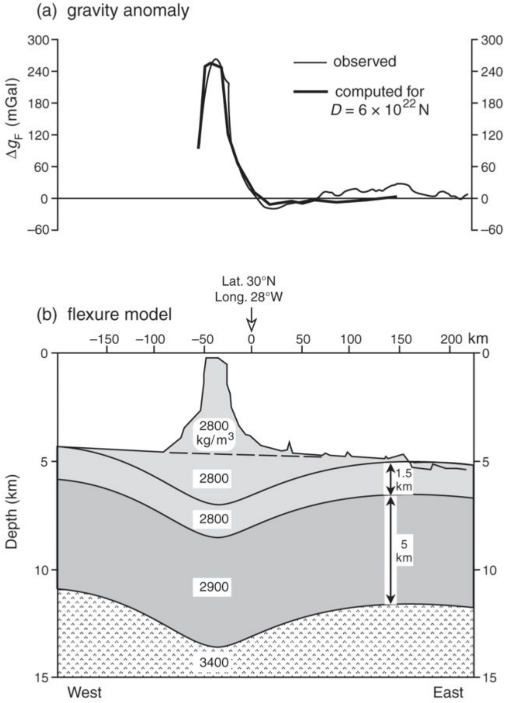


Fig. 5.14 Comparison of (a) the observed free-air gravity anomaly profile across the Great Meteor seamount with (b) the anomaly computed for a lithospheric flexure model of isostatic compensation.

Source: A.B. Watts, J.R. Cochran and G. Selzer, Gravity anomalies and flexure of the lithosphere: a three-dimensional study of the Great Meteor seamount, Northeast Atlantic, *J. Geophys. Res.* 80, 1391–1398, 1975.

5.4.2 Lithospheric Flexure at a Subduction Zone

The bathymetry of an oceanic plate at a subduction zone is typified by an oceanic trench, which can be many kilometers deep (Fig. 5.15a). Seaward of the trench axis the plate develops a small upward bulge (the outer rise), which can extend for 100–150 km away from the trench and reach heights of several hundred meters. The lithospheric plate bends sharply downward in the subduction zone. This bending can also be modeled with a thin elastic plate.

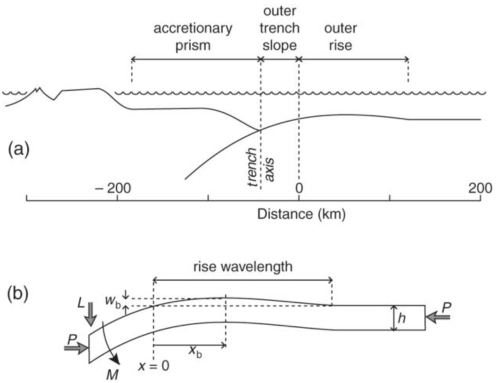


Fig. 5.15 (a) Schematic structural cross-section at a subduction zone. Source: J.G. Caldwell and D.L. Turcotte, Dependence of the thickness of the elastic oceanic lithosphere on age, *J. Geophys. Res.* 84, 7572–7576, 1979. (b) The corresponding thin-plate model.

Source: D.L. Turcotte, D.C. McAdoo and J.G. Caldwell, An elastic-perfectly plastic analysis of the bending of the lithosphere at a trench, *Tectonophysics* 47, 193–205, 1978.

In the model, a horizontal force P pushes a plate of thickness h toward the subduction zone, the leading edge carries a vertical load L, and the plate is bent by a bending moment M (Fig. 5.15b). The horizontal force P is negligible in comparison to the effects of M and L. The vertical deflection of the plate must satisfy Eq. (5.47), with the same parameters as in the previous example. Choosing the origin to be the point nearest the trench where the deflection is zero simplifies the form of the solution, which is

$$w = Ae^{-x/\alpha} \sin\frac{x}{\alpha}, \tag{5.51}$$

where A is a constant and α is the flexural parameter as in Eq. (5.49). The value of the constant A is found from the position x_b of the forebulge, where $dw/dx = 0$:

$$\frac{dw}{dx} = A\left(-\frac{1}{\alpha}e^{-x/\alpha}\sin\frac{x}{\alpha} + \frac{1}{\alpha}e^{-x/\alpha}\cos\frac{x}{\alpha}\right) = 0, \tag{5.52}$$

from which

$$x_b = \frac{\pi}{4}\alpha \qquad \text{and} \qquad A = w_b\sqrt{2}e^{\pi/4}. \tag{5.53}$$

It is convenient to normalize the horizontal distance and vertical displacement: writing $x' = x/x_b$ and $w' = w/w_b$, the generalized equation for the elastic bending at an oceanic trench is obtained:

$$w' = \sqrt{2}\,\sin\left(\frac{\pi}{4}x'\right)\exp\left[\frac{\pi}{4}(1 - x')\right]. \tag{5.54}$$

The theoretical deflection of oceanic lithosphere at a subduction zone obtained from the elastic bending model agrees well with the observed bathymetry on a profile across the Marianas oceanic trench (Fig. 5.16a). The calculated thickness of the elastic lithosphere is about 20–30 km and the flexural rigidity is around 10^{23} N m. However, at some trenches the assumption of a completely elastic upper lithosphere is evidently inappropriate. At the Tonga trench the model curve deviates from the observed bathymetry inside the trench

(Fig. 5.16b). It is likely that the elastic limit is exceeded at parts of the plate where the curvature is high. These regions may yield, leading to a reduction in the effective rigidity of the plate. This effect can be taken into account by assuming that the inelastic deformation is plastic. The deflection calculated within the trench for an elastoplastic model agrees well with the observed bathymetry.

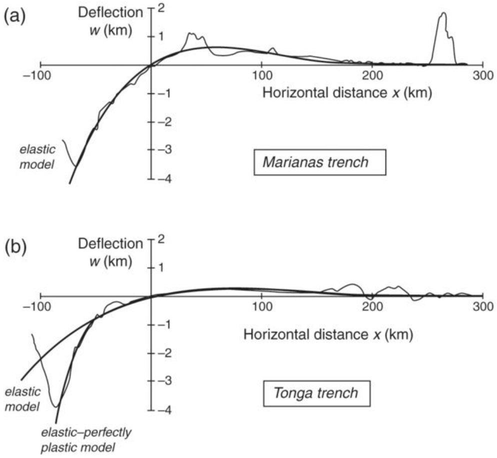


Fig. 5.16 Observed (*solid*) and theoretical (*dashed*) bathymetric profiles at ocean trenches. (a) Elastic flexure of the lithosphere is seen at the Marianas trench. (b) The flexure at the Tonga trench is best explained by an elastic–perfectly plastic.

Source: D.L. Turcotte, D.C. McAdoo and J.G. Caldwell, An elastic-perfectly plastic analysis of the bending of the lithosphere at a trench, *Tectonophysics* 47, 193–205, 1978.

5.4.3 Thickness of Oceanic Lithosphere

Estimates of the elastic thickness of oceanic lithosphere based on the bending plate model can be compared with estimates based on the observation of seismic waves. Surface waves with periods above 30 s penetrate into the upper mantle. The analysis of their frequency-dependent velocity provides estimates of lithospheric thickness, which range from around 50 km to nearly 100 km beneath the oceans. Observations of reflected waves suggest that the boundary between oceanic lithosphere and the underlying asthenosphere is relatively sharp, occupying less than 10 km. This indicates that the lithosphere–asthenosphere boundary may be related to a change in chemical composition or the sudden presence of partial melt within the asthenosphere.

The lithospheric thicknesses obtained from elastic modeling of bending caused by seamounts and island chains or at subduction zones are only one-third to one-half of the seismic thickness. On one hand, the discrepancy suggests that the elastic plate model may be too simple. It ignores viscous effects that become relevant over long timescales, as well as the thermal and compositional layering of the lithosphere. On the other hand, plate bending and seismic wave propagation are controlled by different physical parameters, namely flexural rigidity and wave speed. Changes in these parameters with depth depend on a wide range of factors (composition, temperature, deformation mechanism, presence of fluids), and need not coincide spatially. As a consequence, we distinguish the *elastic lithosphere* inferred from the bending plate model over long timescales, and the *seismic lithosphere,* inferred from the depth-dependent propagation speed of seismic waves over short timescales.

5.5 Mantle Viscosity

As illustrated by the transition from brittle to ductile behavior (Section 5.3.4), the Earth's rheology changes with depth. The upper part of the lithosphere behaves elastically. When stress is sufficiently small to prevent failure, it has a largely reversible response to both short-term and long-term loads.

The deeper part of the lithosphere does not behave purely elastically. Although it has an elastic response to abrupt stress changes – induced, for example, by earthquakes – it reacts to long-lasting stress by ductile flow. This kind of rheological behavior also characterizes the asthenosphere and the deeper mantle. Flow takes place with a strain rate that is proportional to the stress or a power thereof. In the simplest case, the deformation occurs by Newtonian flow governed by a viscosity coefficient η, whose dimensions (Pa s) express the time-dependent nature of the process.

Under a surface load, such as an ice-sheet, the elastic lithosphere is pushed down into the viscous mantle (Fig. 5.17a). This causes an outflow of mantle material away from the depressed region. When the ice-sheet melts, removing the load, hydrostatic equilibrium is restored and there is a return flow of the viscous mantle material (Fig. 5.17b). Thus, in modeling the time-dependent reaction of the Earth to a surface load at least two, usually three, layers must be taken into account. The top layer is an elastic lithosphere up to 100 km thick; it has practically infinite viscosity (i.e., it does not flow) and a flexural rigidity D around 5×10^{24} N m. Beneath it lies a low-viscosity "channel" 75–250 km thick that includes the asthenosphere, with a low viscosity of typically 10^{19}–10^{20} Pa s. The deeper mantle, which makes up the third layer, has a higher viscosity around 10^{21} Pa s.

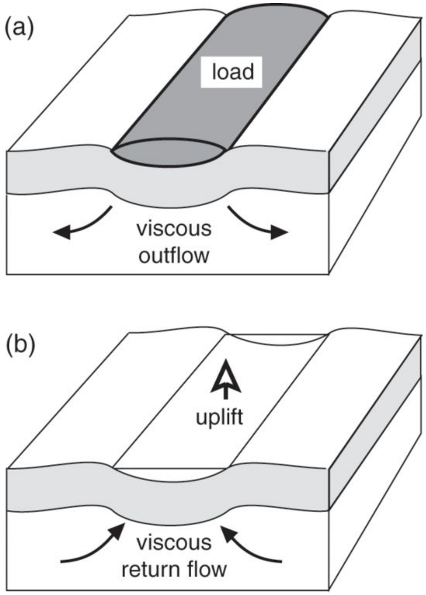


Fig. 5.17 (a) Depression of the lithosphere due to a surface load (e.g., an ice-sheet) and accompanying viscous outflow in the underlying mantle. (b) Return flow in the mantle and surface uplift after removal of the load.

Restoration of the surface after removal of a load is accompanied by uplift, which can be expressed well by a simple exponential relaxation equation. If the initial depression of the surface is w_0, the deflection $w(t)$ after time t is given by

$$w(t) = w_0 e^{-t/\tau}. \tag{5.55}$$

Here, τ is the relaxation time, which is related to the mantle viscosity η by

$$\tau = \frac{4\pi}{\rho_m g \lambda}\eta, \tag{5.56}$$

where ρ_m is the mantle density, g is gravity at the depth of the flow, and λ is the wavelength of the depression, a dimension appropriate to the scale of the load (as will be seen in the following examples). A test of these relationships requires data that give surface elevations in the past. These data come from analyses of sea-level changes, present elevations of previous shorelines, and directly observed rates of uplift. The ancient horizons have been dated by radiometric methods as well as by sedimentary methods such as varve chronology, which consists of counting the annually deposited pairs of silt and clay layers in laminated sediments. A good example of the exponential restoration of a depressed region is the history of uplift in central Fennoscandia since the end of the last ice age some 10,000 years ago (Fig. 5.18). If it is assumed that about 30 m of uplift still remain, the observed uplift agrees well with Eq. (5.55) and gives a relaxation time of 4400 years.

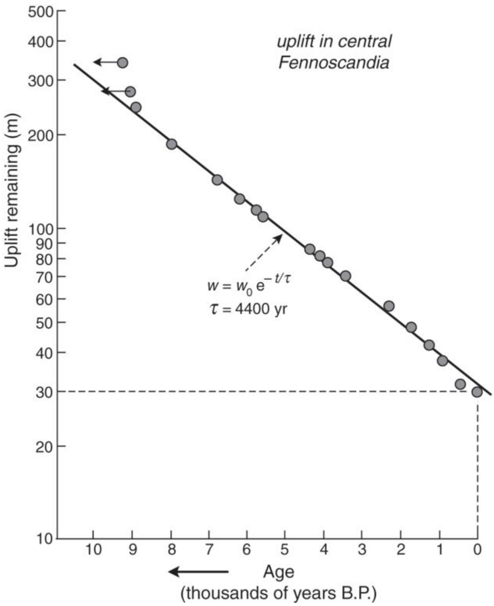


Fig. 5.18 Uplift in central Fennoscandia since the end of the last ice age illustrates exponential viscous relaxation with a time constant of 4400 years.

Source: L.M. Cathles, *The Viscosity of the Earth's Mantle*, 388 pp., Princeton, N.J.: Princeton University Press, 1975.

An important factor in modeling uplift is whether the viscous response is caused by the mantle as a whole, or whether it is confined to a low-viscosity layer (or “channel”) beneath the lithosphere. Seismic shear-wave velocities are reduced in a low-velocity channel, the thickness of which varies from one locality to another between a few tens to nearly 200 km. Interpreted as the result of softening or partial melting due to temperatures near the melting point, this seismic low-velocity channel is called the asthenosphere. It is represented by a distinct layer where viscosity is at least 25 times less than in the deeper mantle.

The larger the ice-sheet (or other type of surface load), the deeper the effects reach into the mantle. By studying the uplift following removal of different loads, information is obtained about the viscosity at different depths in the mantle.

5.5.1 Viscosity of the Upper Mantle

About 18,000–20,000 years ago Lake Bonneville, the predecessor of the present Great Salt Lake in Utah, USA, had a radius of about 95 km and an estimated depth of about 305 m. The water mass depressed the lithosphere, which was later uplifted by isostatic restoration after the lake drained and dried up. Observations of the present heights of ancient shorelines show that the central part of the lake has been uplifted by about 65 m. Two parameters are involved in the process: the flexural rigidity of the lithosphere and the viscosity of the mantle beneath. The elastic response of the lithosphere is estimated from the geometry of the depression that would be produced in isostatic equilibrium. The maximum flexural rigidity that would allow a 65 m deflection under a 305 m water load is found to be about 5×10^{23} N m.

The surface load can be modeled as a heavy vertical right cylinder with radius r, which pushes itself into the soft mantle. The underlying viscous material is forced aside so that the central depression is surrounded by a circular uplifted “bulge.” After removal of the load, restorative uplift takes place in the central depression, the peripheral bulge subsides, and the contour of zero-uplift migrates outward. The wavelength λ of the depression caused by a load with this geometry has been found to be about 2.6 times the diameter of the cylindrical load. In the case of Lake Bonneville, $2r = 192$ km, so λ is about 500 km. The mantle viscosity is obtained by assuming that the response time of the lithosphere was short enough to track the loading history quite closely. This implies that the viscous relaxation time t must have been 4000 years or less. Substitution of these values for λ and τ in Eq. (5.56) suggests a maximum mantle viscosity η of about 2×10^{20} Pa s in the top 250 km of the mantle. A lower value of η would require a thinner low-viscosity channel beneath the lithosphere.

The Fennoscandian uplift can be treated in the same way (Fig. 5.19), but the weight and lateral expanse of the load were much larger. The ice-cap is estimated to have been about 1100 m thick and to have covered most of Norway, Sweden, and Finland. Although the load was therefore somewhat elongate, it is possible to model it satisfactorily by a vertical cylinder of radius $r = 550$ km centered on the northern Gulf of Bothnia (Fig. 5.20). The load was applied for about 20,000 years before being removed 10,000 years ago. This caused an initial depression of about 300 m, which has subsequently been relaxing (Fig. 5.19). The uplift rates allow upper mantle viscosities to be estimated. The data are compatible with different models of mantle structure, two of which are compared in Fig. 5.20. Each model has an elastic lithosphere with a flexural rigidity of 5×10^{24} N m underlain by a low-viscosity channel and the rest of the mantle. The first model has a 75 km thick channel ($\eta = 4 \times 10^{19}$ Pa s) over a viscous mantle ($\eta = 10^{21}$ Pa s). The alternative model has a 100 km thick channel (viscosity coefficient $\eta = 1.3 \times 10^{19}$ Pa s) over a rigid mantle. Both models yield uplift rates that agree quite well with the observed uplift rates.

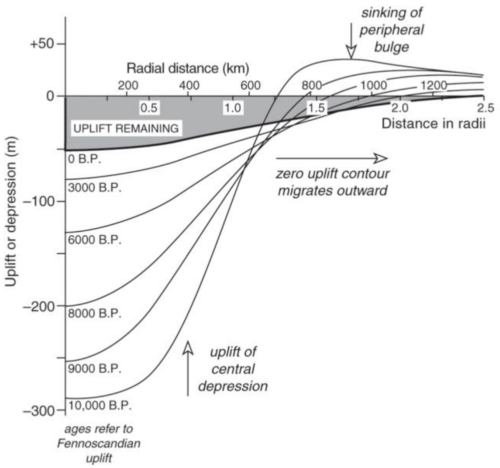
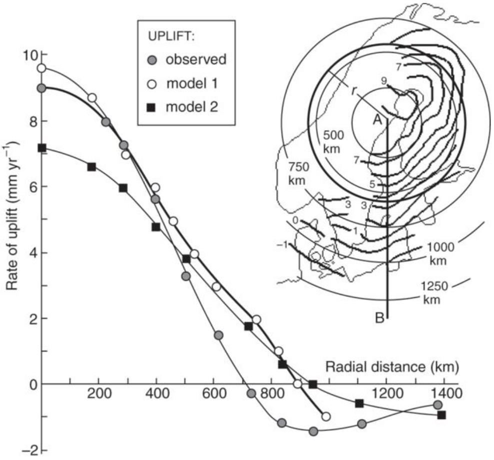


Fig. 5.19 Model calculations of the relaxation of the deformation caused by the Fennoscandian ice-sheet following its disappearance 10,000 years ago.

Source: L.M. Cathles, *The Viscosity of the Earth's Mantle*, 388 pp., Princeton, N.J.: Princeton University Press, 1975.



Model 1:
lithosphere: $D = 5 \times 10^{24}$ N m
low viscosity channel:
thickness = 75 km
$\eta = 4 \times 10^{19}$ Pa s
mantle: $\eta = 10^{21}$ Pa s

Model 2:
lithosphere: $D = 5 \times 10^{24}$ N m
low viscosity channel:
thickness = 100 km
$\eta = 1.3 \times 10^{19}$ Pa s

Fig. 5.20 Comparison of Fennoscandian uplift rates interpreted along profile AB (inset) with uplift rates calculated for two different models of mantle viscosity, assuming the ice-sheet can be represented by a vertical right cylindrical load centered on the northern Gulf of Bothnia.

Source: L.M. Cathles, *The Viscosity of the Earth’s Mantle*, 388 pp., Princeton, N.J.: Princeton University Press, 1975.

5.5.2 Viscosity of the Lower Mantle

Geologists have developed a coherent picture of the last glacial stage, the Wisconsin, during which much of North America was covered by an ice-sheet over 3500 m thick. The ice persisted for about 20,000 years and melted about 10,000 years ago. It caused a surface depression of about 600 m. The subsequent history of uplift in the James Bay area near the center of the feature has been reconstructed using geological indicators and dated by the radiocarbon method (see Section 8.4.1). For modeling purposes the ice-sheet can be represented as a right cylindrical load with radius $r = 1650$ km (Fig. 5.21, *inset*). A load as large as this affects the deep mantle. The central uplift following removal of the load has been calculated for several Earth models. Each model has elastic parameters and density distribution obtained from seismic velocities, including a central dense core. The models differ from each other in the number of viscous layers in the mantle and the amount by which the density gradient departs from the adiabatic gradient (Table 5.1). The curvature of the observed uplift curve only fits models in which the viscosity of the lower mantle is around 10^{21} Pa s (Fig. 5.21). A highly viscous lower mantle ($\eta = 10^{23}$ Pa s, model #4) is incompatible with the observed uplift history. The best fit is obtained with models #1 or #5. Each has an adiabatic density gradient, but model #1 has a uniform lower mantle with viscosity 10^{21} Pa s, and model #5 has η increasing from 10^{21} Pa s below the lithosphere to 3×10^{21} Pa s just above the core.

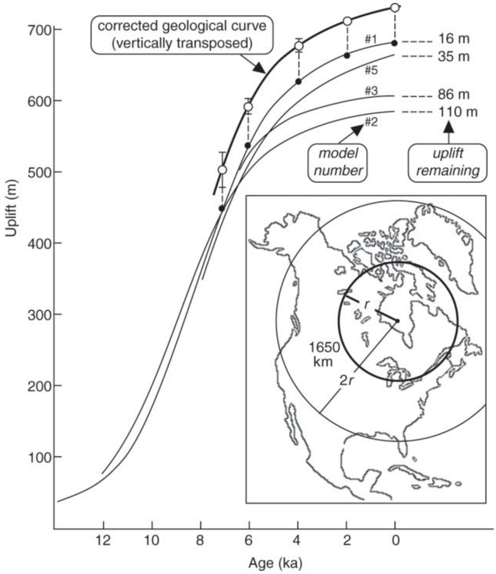


Fig. 5.21 Comparison of uplift history in the James Bay area with predicted uplifts for various Earth models after disappearance of the Wisconsin ice-sheet over North America, represented as a vertical right cylindrical load with radius 1650 km as in the inset. For details of model parameters, see Table 5.1.

Source: L.M. Cathles, *The Viscosity of the Earth's Mantle*, 388 pp., Princeton, N.J.: Princeton University Press, 1975.

Table 5.1 Parameters of Earth models used in computing uplift rates. All models are for an elastic Earth with a dense core.

Model	Density gradient	Viscosity (10^{21} Pa s)	Depth interval
1	Adiabatic	1	Entire mantle
2	Adiabatic, except 335–635 km	1	Entire mantle
3	Adiabatic, except 335–635 km	0.1	0–335 km
		1	335 km to core
4	Adiabatic	1	0–985 km
		100	985 km to core
5	Adiabatic	1	0–985 km
		2	985–2185 km

3	2185 km to core

Source: after Cathles, 1975.

The viscoelastic properties of the Earth's interior influence the Earth's rotation, causing changes in the position of the instantaneous rotation axis. The motion of the axis can be traced, for example, by repeated measurements in which the zenith is located by photographing the stars vertically above an observatory. The development of very long baseline interferometry (VLBI) and other space geodetic methods (Section 3.5) has provided alternative accurate methods of observing the location of the instantaneous rotation axis. The observations reveal systematic movements of the rotation axis relative to the axis of figure (Fig. 5.22). Decomposed into components along the Greenwich meridian (x-axis) and the 90°W meridian (y-axis), the polar motion exhibits a fluctuation with cyclically varying amplitude superposed on a linear trend. The amplitude modulation has a period of approximately seven years and is due to the interference of the 12-month annual wobble and the 14-month Chandler wobble. The linear trend represents a slow drift of the pole toward northern Canada at a rate of 0.95 degrees per million years. It is due to the melting of the Fennoscandian and Laurentide ice-sheets. The subsequent uplift constitutes a redistribution of mass that causes modifications to the Earth's moments and products of inertia, thus affecting the rotation.

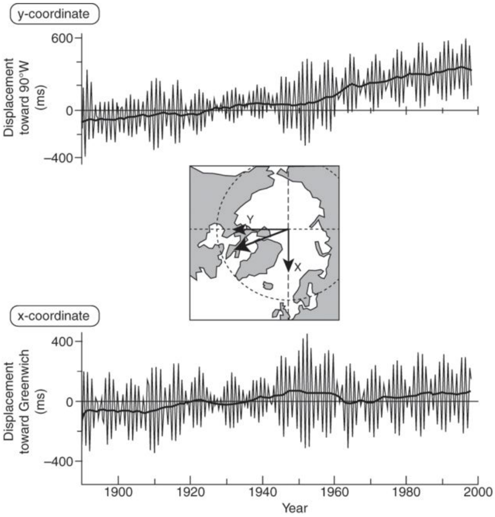


Fig. 5.22 Motion of the instantaneous rotation axis from 1890 to 2000 relative to axes defined in the inset (data source: Gambis and Bizouard, 2014). The thick line is the long-term trend in the position of the rotation pole.

Source: International Earth Rotation and Reference Systems Service.

The observed polar drift can be modeled with different viscoelastic Earth structures. The models assume a 120 km thick elastic lithosphere and take into account different viscosities in the layers bounded by the seismic discontinuities at 410 km and 660 km depths (Sections 7.3 and 7.4.6.3), and the core–mantle boundary. An Earth that is homogeneous below the lithosphere (without a core) is found to give imperceptible drift. Inclusion of the core, with a density jump across the core–mantle boundary and assuming the mantle viscosity to be around 1×10^{21} Pa s, gives a drift that is perceptible but much slower than that observed. Introduction of a density change at the 660 km discontinuity increases the drift markedly; the 400 km discontinuity does not noticeably change the drift further. The optimum model has an upper mantle viscosity of about 1×10^{21} Pa s and a lower mantle viscosity of about 3×10^{21} Pa s. The model satisfies both the rate of drift and its direction (Fig. 5.22). The viscosities are comparable to values found by modeling post-glacial uplift (Table 5.1, model #5).

5.6 *Suggestions for Further Reading*

Intermediate and Advanced Level

Cathles, L. M. 1975. *The Viscosity of the Earth's Mantle*. Princeton, NJ: Princeton University Press.

Fowler, C. M. R. 2004. *The Solid Earth: An Introduction to Global Geophysics* (2nd ed.). Cambridge: Cambridge University Press.

Karato, S.-I. 2012. *Deformation of Earth Materials*. Cambridge: Cambridge University Press.

Kennett, B. L. N. and Bunge, H.-P. 2018. *Geophysical Continua*. Cambridge: Cambridge University Press.

Ranalli, G. 1995. *Rheology of the Earth*. London: Chapman and Hall.

Turcotte, D. L. and Schubert, G. 2002. *Geodynamics* (2nd ed.). Cambridge: Cambridge University Press.

Watts, A. B. 2001. *Isostasy and Flexure of the Lithosphere*. Cambridge: Cambridge University Press.

5.7 Review Questions

1. What are the defining characteristics of elastic deformation?

2. What is Hooke's law, and what are the limits of its validity?

3. What is the meaning of the diagonal and the off-diagonal elements of the stress matrix, respectively?

4. Describe in words the physical meaning of the shear and the bulk modulus, respectively.

5. How is elastic anisotropy related to deformation in the Earth?

6. How can it be that solid material flows? Over which timescales can such solid-state flow be observed?

7. Explain, in your own words, the concepts of dislocation creep, diffusion creep, and thermal activation of a deformation mechanism.

8. Describe why and how the deformation of real materials deviates from the idealized elastic and viscous deformation styles.

9. What is the Maxwell time of a viscoelastic material? How does it relate to the elastic and the viscous properties of the material?

10. Explain the difference between solid-state viscous flow and plastic deformation.

11. What controls the transition between brittle and ductile deformation in oceanic and continental crust?

12. Explain the physical meaning of the flexural rigidity, D.

13. Why does oceanic lithosphere bend upwards near a subduction zone?

14. What is post-glacial uplift? Why does it occur? In which regions of the Earth can post-glacial uplift currently be observed?

5.8 Exercises

1. Propose an experiment to measure Poisson's ratio of a rock sample in the laboratory.

2. A cube of some material is submerged in water, where it experiences hydrostatic pressure of $p = 10$ kPa. The pressure results in the strains $\varepsilon_{xx} = \varepsilon_{yy} = \varepsilon_{zz} = -0.01$ and $\varepsilon_{xy} = \varepsilon_{xz} = \varepsilon_{yz} = 0$.

(a) Write down the strain matrix.

(b) What is the dilatation (fractional volume change) that corresponds to this strain?

(c) What is the bulk modulus of the material?

3. Express the Lamé parameter λ in terms of Poisson's ratio and the shear modulus. Can λ be negative?

4. Consider a water layer with thickness 1 m, e.g., a shallow river. The flow velocity is 0 km s^{-1} at the bottom and 5 km s^{-1} at the surface. The viscosity of water, at 10°C temperature, is 1.3 mPa s. What is the shear stress within the water layer?

5. In a rock deformation laboratory, you expose a sample to a fixed stress σ. As the experiment proceeds, you increase the temperature (in kelvin) by 10%. Knowing that the activation energy of your sample is $E_a = 500$ kJ mol^{-1}, how do you expect the deformation rate to increase as a result of heating?

6. In your laboratory, you also investigate samples of fine-grained sedimentary rock. Which kind of experiment would you conduct in order to reveal the creep mechanism by which the samples deform? Argue on the basis of the constitutive equations for diffusion and dislocation creep.

7. A viscoelastic material is subjected to a constant stress σ, as a consequence of which it deforms gradually toward a final strain ε_m. The measured Maxwell time is τ. Derive an equation for the viscosity of the material in terms of σ, ε_m, and τ.

8. Consider the deformation of the ocean floor in response to loading in the form of an island. How is the wavelength of the deformation related to the thickness of the lithospheric plate? Argue in terms of equations and draw a diagram of the thickness–wavelength relation.

9. Post-glacial uplift occurs nearly exponentially over time (Eq. 5.55). Based on the dating of sedimentary rocks, you know that sample 1, located at a present elevation of 10 m, formed around 1000 years ago. Sample 2, at 5 m elevation, is 500 years old.

- **(a)** What is the corresponding relaxation time of the upper mantle?
- **(b)** The wavelength of the glacial depression is around 1000 km. Provide an estimate of the upper mantle viscosity.

5.9 Computational Exercises

Jupyter notebook computational exercises can be found online for this chapter at www.cambridge.org/FoG3ed (Creep mechanisms).

6

Seismology

Preview

Just as air acts as the carrier of sound waves, the Earth supports the propagation of seismic waves, which are excited, for example, by earthquakes, eruptions and explosions. In human history, high-amplitude seismic waves and related events such as tsunamis have brought more destruction than any other natural phenomenon. However, seismic waves also transmit information about their source and about the medium through which they propagate. In this chapter, we introduce the basic elements of seismic wave propagation, starting with a description of different wave types that are distinguished by their propagation velocities, polarization directions, and the regions of the Earth's interior through which they travel. We will learn how seismic waves interact with heterogeneities, i.e., how they reflect, refract, and convert into each other. This will allow us to decipher the wealth of information transmitted by these waves in order to construct images of the deep Earth's interior.

6.1 Introduction

Seismology is a venerable science with a long history. The Chinese scientist Chang Heng is credited with the invention in AD 132, nearly 2000 years ago, of the first functional seismoscope, an ingenious device of elegant construction and beautiful design that registered the arrival of seismic waves and enabled the observer to infer the direction from which they came. The origins of earthquakes were not at all understood. For centuries these fearsome events were attributed to supernatural powers. The accompanying destruction and loss of life were often understood in superstitious terms and interpreted as punishment inflicted by the gods on a sinful society. Although early astronomers and philosophers sought to explain earthquakes as natural phenomena unrelated to spiritual factors, the belief that earthquakes were an expression of divine anger prevailed until the advent of the Age of Reason in the eighteenth century. The path to a logical understanding of natural phenomena was laid in the seventeenth century by the systematic observations of scientists like Galileo, the discovery and statement of physical laws by Newton, and the development of rational thought by contemporary philosophers.

In addition to the development of the techniques of scientific observation, an understanding of the laws of elasticity and the limited strength of materials was necessary before seismology could progress as a science. In a pioneering study, Galileo in 1638 described the response of a beam to loading, and in 1660 Hooke established the law of the spring. However, another 150 years were to pass before the generalized equations of elasticity were set down by Navier. During the early decades of the nineteenth century, Cauchy and Poisson completed the foundations of modern elasticity theory.

Elastic waves emitted by earthquakes propagate through the entire Earth. Just as electromagnetic waves in the form of X-rays carry information on the interior of the human body, seismic waves can be used to constrain the internal structure of the Earth. Replacing earthquakes by artificial, controlled explosions, the techniques of refraction and reflection seismology were developed in the search for hydrocarbon and geothermal resources, or suitable construction sites for dams, bridges, and other large infrastructure and buildings. These active source methods can also be applied successfully to the resolution of detailed crustal structure under continents and oceans. The development of powerful computer technology enabled refinements in earthquake location and in the determination of travel times of seismic waves. These advances led to the modern field of seismic tomography, a powerful technique for revealing regions of the Earth's interior that have anomalous seismic velocities. In the field of earthquake seismology, the need to protect populations and man-made structures has resulted in the investment of considerable effort in the study of earthquake hazard and the development of construction codes to reduce earthquake damage.

What distinguishes seismology from other wave-related sciences is the remarkable proliferation of seismic waves inside the Earth, exemplified by the recordings (called *seismograms*) shown in Fig. 6.1. Waves with different propagation velocities and polarizations travel through the interior of the Earth and along its surface. They reflect at the surface and at discontinuities inside the Earth, where they convert into each other.

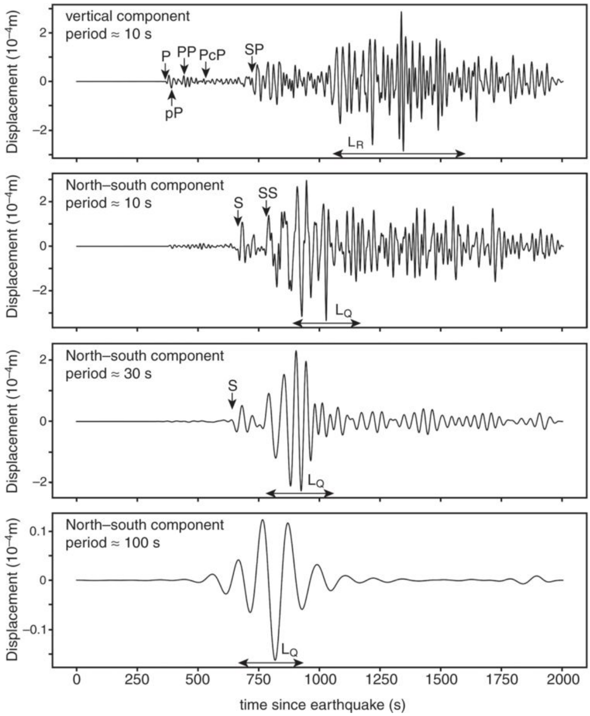


Fig. 6.1 Seismograms recorded at a seismological station in the Swiss Alps documenting a large, magnitude 7.3 earthquake that occurred on November 12, 2017 in the Zagros mountains in Iran. (a) Vertical component of ground displacement at a period of around 10 s. (b)–(d) North–south component of ground displacement at periods of around 10 s, 30 s, and 100 s. A selection of seismic wave types, explained in Section 6.2, is indicated. Each of these waves propagates through different parts of the Earth.

To appreciate how seismologists have deciphered the structure of the Earth's interior from this wealth of information, it is necessary to understand what types of seismic waves can be generated by an earthquake or man-made source. Though material near the seismic source suffers permanent deformation, the propagation of seismic waves outside the source is governed by the principles of elasticity theory (Section 5.1), and therefore by the distributions of physical properties such as density and elastic moduli.

6.2 Seismic Waves

6.2.1 Introduction

The propagation of seismic waves through the Earth can be triggered by a wide range of sources, including earthquakes, man-made explosions, volcanic eruptions, landslides, and meteorite impacts. Close to the source, the medium is often permanently destroyed, which means that the deformation is non-recoverable. However, at some distance from the source, deformation amplitudes are small and the medium deforms (an)elastically. The particles of the medium carry out harmonic oscillations, and energy is transmitted over large distances in the form of waves.

Seismic waves can be classified into two main types. When energy is released near the surface (Fig. 6.2), part of it propagates through the body of the medium as *body waves*. The remaining energy spreads out over the surface as *surface waves*, analogous to the ripples on the surface of a pool of water into which a stone has been thrown.

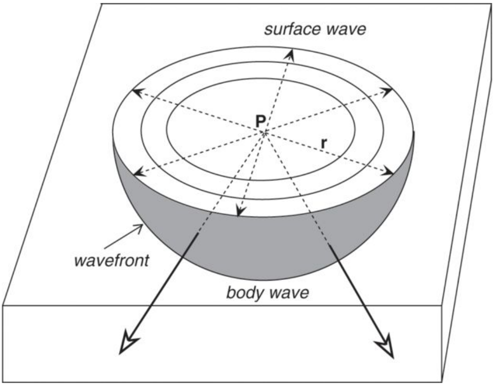


Fig. 6.2 Propagation of seismic waves from a point source P near the surface. The disturbance travels as a body wave through the medium and as a surface wave along the free surface. When the medium is homogeneous, wavefronts are perfect hemispheres for body waves and perfect circles for surface waves.

6.2.2 Seismic Body Waves

When a body wave reaches a distance r from its source in a homogeneous medium, the *wavefront* (defined as the surface in which all particles vibrate with the same phase) has a spherical shape, and the wave is called a *spherical wave*. As the distance from the source increases, the curvature of the spherical wavefront decreases. At great distances from the source the wavefront is so flat that it can locally be considered to be a plane and the seismic wave is called a *plane wave*. The direction perpendicular to the wavefront is called the seismic *ray path*. The description of the harmonic motion in plane waves is simpler than for spherical waves, because for plane waves we can use orthogonal Cartesian coordinates.

6.2.2.1 Compressional Waves

Let Cartesian reference axes be defined such that the x-axis is parallel to the direction of propagation of the plane wave; the y- and z-axes then lie in the plane of the wavefront (Fig. 6.3). Vibrations of the medium have components parallel to each of the axes. In the x-direction the particle motion is back and forward parallel to the direction of propagation. This results in the medium being alternately stretched and condensed in this direction (Fig. 6.4a). This harmonic motion produces a body wave that is transmitted as a sequence of rarefactions and condensations parallel to the x-axis.

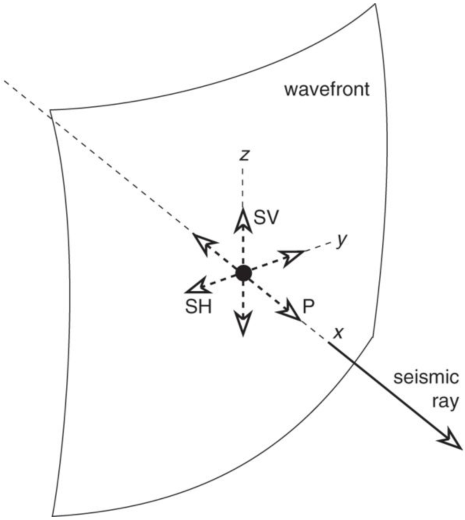

wavefront
z
SV
y
SH
P
x
seismic ray

Fig. 6.3 Representation of a vibration as components parallel to three orthogonal reference axes. Particle motion in the *x*-direction is back and forth parallel to the direction of propagation, corresponding to the P-wave. Vibrations along the *y*- and *z*-axes are in the plane of the wavefront and normal to the direction of propagation. The *z*-vibration in a vertical plane corresponds to the SV-wave; the *y*-vibration is horizontal and corresponds to the SH-wave.

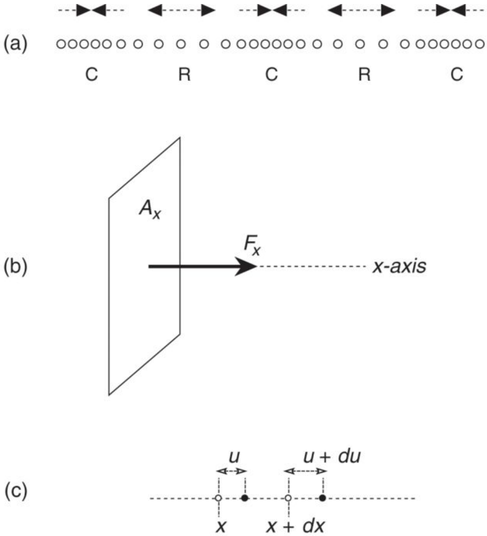

(a)
C
R
C
R
C
(b)
A_x
F_x
x-axis
(c)
u
u + du
x
x + dx

Fig. 6.4 (a) The particle motion in a one-dimensional P-wave transmits energy as a sequence of rarefactions (R) and condensations (C) parallel to the x-axis. (b) Within the wavefront the component of force F_x in the x-direction of propagation is distributed over an element of area A_x normal to the x-axis. (c) A particle at position x experiences a longitudinal displacement u in the x-direction, while at the nearby position $x + \mathrm{d}x$ the corresponding displacement is $u + \mathrm{d}u$.

Consider the disturbance of the medium shown in Fig. 6.4b. The area of the wavefront normal to the x-direction is A_x, and the wave propagation is treated as one-dimensional. At an arbitrary position x (Fig. 6.4c), the passage of the wave produces a displacement u and a force F_x in the x-direction. At the position $x+\mathrm{d}x$ the displacement is $u + \mathrm{d}u$ and the force is $F_x + \mathrm{d}F_x$. Here, $\mathrm{d}x$ is the infinitesimal length of a small volume element which has mass $\rho \mathrm{d}x\, A_x$. The net force acting on this element in the x-direction is given by

$$(F_x + \mathrm{d}F_x) - F_x = \frac{\partial F_x}{\partial x}\mathrm{d}x. \tag{6.1}$$

The force F_x is caused by the stress element σ_{xx} acting on the area A_x, and is equal to $\sigma_{xx} A_x$. This allows us to write the one-dimensional equation of motion (Newton's second law),

$$(\rho \mathrm{d}x A_x)\frac{\partial^2 u}{\partial t^2} = \mathrm{d}x A_x \frac{\partial \sigma_{xx}}{\partial x}, \tag{6.2}$$

which describes the balance between inertial (acceleration-related) forces on the left-hand side, and elastic (deformation-related) forces on the right-hand side. The definitions of Young's modulus, E (Eq. 5.15), and the normal strain ε_{xx} (Eq. 5.4) give, for a one-dimensional deformation,

$$\sigma_{xx} = E\varepsilon_{xx} = E\frac{\partial u}{\partial x}. \tag{6.3}$$

Substitution of Eq. (6.3) into Eq. (6.2) gives the one-dimensional wave equation,

$$\frac{\partial^2 u}{\partial t^2} = \alpha^2 \frac{\partial^2 u}{\partial x^2}, \tag{6.4}$$

where α is given by

$$\alpha = \sqrt{\frac{E}{\rho}}. \tag{6.5}$$

The parameter α has the dimensions of a velocity, m s^{-1}. As we will see in Section 6.2.2.3, α is indeed the propagation velocity of a compressional wave.

A one-dimensional wave is rather restrictive. It represents the stretching and compressing in the x-direction as effects that are independent of what happens in the y- and z-directions. In an elastic solid, the elastic strains in any direction are coupled to the strains in transverse directions by Poisson's ratio for the medium. A three-dimensional analysis is given in Appendix A that takes into account the simultaneous changes perpendicular to the direction of propagation. In this case the area A_x can no longer be considered constant, and the longitudinal (or compressional) body wave passes through a medium as harmonic volume changes, or *dilatations* θ (Eqs. 5.7 and 5.8). The equation of the compressional wave in the x-direction for the three-dimensional case is

$$\frac{\partial^2 \theta}{\partial t^2} = \alpha^2 \frac{\partial^2 \theta}{\partial x^2},$$
(6.6)

where the compressional wave velocity α is now given by

$$\alpha = \sqrt{\frac{\lambda + 2\mu}{\rho}} = \sqrt{\frac{K + \frac{4}{3}\mu}{\rho}}.$$
(6.7)

The longitudinal wave is the fastest of all seismic waves. When an earthquake occurs, this wave is the first to arrive at a recording station, and is therefore called the *primary wave*, or *P-wave*. Equation (6.7) shows that P-waves can travel through solids, liquids, and gases, all of which are compressible ($K \neq 0$). Liquids and gases do not allow shear. Consequently, $\mu = 0$, and the compressional wave velocity in a liquid or gas is given by

$$\alpha = \sqrt{\frac{K}{\rho}}.$$
(6.8)

Several P-waves are marked in the seismogram example of Fig. 6.1a. Since the Earth is not homogeneous, there are multiple P-waves, labeled for instance P, pP, or PcP. These different P-waves travel through the Earth along different paths and therefore arrive at different times. All P-waves travel as coherent wave packets that are distinguishable from each other. These individual wave packets are called *seismic phases*.

6.2.2.2 Transverse Waves

The vibrations along the *y*- and *z*-axes (Fig. 6.3) are parallel to the wavefront and transverse to the direction of propagation. If we wish, we can combine the y- and z-components into a single transverse motion. It is more convenient, however, to analyze the motions in the vertical and horizontal planes separately. Here, we discuss the disturbance in the vertical plane defined by the x- and z-axes; an analogous description applies to the horizontal plane.

The transverse wave motion is akin to that seen when a rope is shaken. Vertical planes move up and down and adjacent elements of the medium experience shape distortions (Fig. 6.5a), changing repeatedly from a rectangle to a parallelogram and back. Adjacent elements of the medium suffer vertical shear.

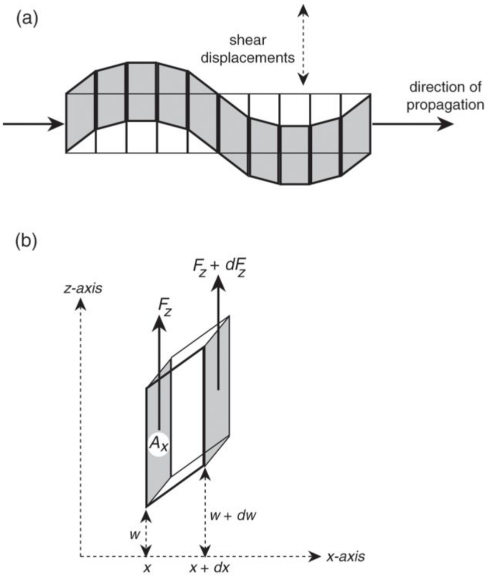


Fig. 6.5 (a) Shear distortion caused by a one-dimensional S-wave. (b) Displacements and forces in the z-direction at the positions x and $x + \mathrm{d}x$ bounding a small sheared element.

Consider the distortion of an element bounded by vertical planes separated by a small horizontal distance $\mathrm{d}x$ (Fig. 6.5b) at an arbitrary horizontal position x. The passage of a wave in the x-direction produces a displacement w and a force F_z in the z-direction. At the position $x + \mathrm{d}x$ the displacement is $w + \mathrm{d}w$ and the force is $F_z + \mathrm{d}F_z$.

The mass of the small volume element bounded by the vertical planes is $\rho \mathrm{d}x\, A_x$, where A_x is the area of the bounding plane. The net force acting on this element in the z-direction is given by

$$(F_z + \mathrm{d}F_z) - F_z = \mathrm{d}F_z = \frac{\partial F_z}{\partial x}\mathrm{d}x. \tag{6.9}$$

The force F_z arises from the shear stress σ_{xz} on the area A_x, and is equal to $\sigma_{xz} A_x$. The equation of motion of the vertically sheared element is

$$(\rho \mathrm{d}x A_x)\frac{\partial^2 w}{\partial t^2} = \mathrm{d}x A_x \frac{\partial \sigma_{xz}}{\partial x}. \tag{6.10}$$

We now have to modify the Lamé expression for Hooke's law and the definition of shear strain so that they apply to the passage of a one-dimensional shear wave in the x-direction. In this case, because the areas of the parallelograms between adjacent vertical planes are equal, there is no volume change. The dilatation θ is zero, and Hooke's law is as given in Eq. (5.16):

$$\sigma_{xz} = 2\varepsilon_{xz}. \tag{6.11}$$

Following the definition of shear–strain components in Eq. (5.12) we have

$$\varepsilon_{xz} = \frac{1}{2}\left(\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z}\right). \tag{6.12}$$

For a one-dimensional shear wave there is no change in the distance $\mathrm{d}x$ between the vertical planes; $\mathrm{d}u$ and $\partial u/\partial z$ are zero and ε_{xz}

is equal to $(\partial w/\partial x)/2$. On substitution into Eq. (6.11) this gives

$$\sigma_{xz} = \mu \frac{\partial w}{\partial x}, \tag{6.13}$$

and on further substitution into Eq. (6.10) and rearrangement of terms we get

$$\frac{\partial^2 w}{\partial t^2} = \beta^2 \frac{\partial^2 w}{\partial x^2}, \tag{6.14}$$

where β is given by

$$\beta = \sqrt{\frac{\mu}{\rho}}. \tag{6.15}$$

Similar to the case of a compressional wave (Section 6.2.2.1), β is the propagation velocity of a shear wave, with units of m s^{-1}. The only elastic property that determines the velocity of the shear wave is the rigidity or shear modulus, μ. In liquids and gases μ is zero and shear waves cannot propagate. In solids, a quick comparison of Eqs. (6.7) and (6.15) gives

$$\alpha^2 - \frac{4}{3}\beta^2 = \frac{K}{\rho}. \tag{6.16}$$

By definition, the bulk modulus K is positive (if it were negative, an increase in confining pressure would cause an increase in volume), and therefore α is always greater than β. Shear waves from an earthquake travel more slowly than P-waves and are recorded as later arrivals. Shear waves are often referred to as *secondary waves* or *S-*

waves.

The shear wave motion within the plane of the wavefront can be resolved into two orthogonal components, one horizontal in the *y*-direction and the other in the vertical *z*-direction (Fig. 6.3). Equation (6.14) describes a one-dimensional shear wave that travels in the *x*-direction, but that has particle displacements (*w*) in the *z*-direction. This vertically polarized wave is called the *SV-wave*. A similar equation describes the shear wave propagating in the *x*-direction with particle displacements (*v*) in the *y*-direction. A shear wave that is polarized horizontally is called an *SH-wave*.

As for the description of longitudinal waves, this treatment of shear wave transmission is somewhat simplified, as it ignores that shear waves may also propagate in directions other than the *x*-direction. A complete three-dimensional treatment is given in Appendix A. The passage of a shear wave involves rotations of volume elements within the plane normal to the ray path, without changing their volume. For this reason, shear waves are also sometimes called rotational (or equivoluminal) waves. The rotation is a vector, Ψ, with *x*-, *y*-, and *z*-components given by

$$\psi_x = \frac{\partial w}{\partial y} - \frac{\partial v}{\partial z} \qquad \psi_y = \frac{\partial u}{\partial z} - \frac{\partial w}{\partial x} \qquad \psi_z = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}. \tag{6.17}$$

A more appropriate equation for the shear wave in the *x*-direction is then

$$\frac{\partial^2 \Psi}{\partial t^2} = \beta^2 \frac{\partial^2 \Psi}{\partial x^2}, \tag{6.18}$$

where β is again the shear-wave velocity as given by Eq. (6.15).

Until now we have chosen the direction of propagation along one of the reference axes so as to simplify the mathematics. If we remove this restriction, additional second-order differentiations with respect to the *y*- and *z*-coordinates must be introduced. The P-wave and S-wave equations become, respectively,

$$\frac{\partial^2 \theta}{\partial t^2} = \alpha^2 \left(\frac{\partial^2 \theta}{\partial x^2} + \frac{\partial^2 \theta}{\partial y^2} + \frac{\partial^2 \theta}{\partial z^2} \right), \tag{6.19}$$

$$\frac{\partial^2 \Psi}{\partial t^2} = \beta^2 \left(\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} + \frac{\partial^2 \Psi}{\partial z^2} \right). \tag{6.20}$$

Several S-wave phases are marked in the seismogram examples in Fig. 6.1. As predicted by theory, S-waves arrive later than P-waves. They also travel as coherent wave packets that can be classified into different S-wave phases.

6.2.2.3 The Solution of the Seismic Wave Equation in a Homogeneous Medium

Two important characteristics of a wave motion are: (1) it transmits energy by means of elastic displacements of the particles of the medium, i.e., there is no net transfer of mass; and (2) the wave pattern repeats itself exactly in both time and space when the medium is homogeneous. The harmonic repetition allows us to express the amplitude variation by a *sine* or *cosine* function. As the wave passes any point, the amplitude of the disturbance is repeated at regular time intervals, *T*, the *period* of the wave. The number of times the amplitude is repeated per second is the *frequency, f*, which is equal to the inverse of the period ($f = 1/T$). At any instant in time, the disturbance in the medium is repeated along the direction of travel at regular distances, λ, the *wavelength* of the wave. During the propagation of a wave in the *x*-direction, the harmonic displacement

(u) of a particle from its mean position can be written as

$$u = A \sin\left[2\pi\left(\frac{x}{\lambda} - \frac{t}{T}\right)\right], \tag{6.21}$$

where A is the amplitude.

The quantity in round brackets is called the *phase* of the wave. Any value of the phase corresponds to a particular amplitude and direction of motion of the particles of the medium. The *wave number* (k), *angular frequency* (ω), and velocity (c) are defined and related by

$$k = \frac{2\pi}{\lambda} \qquad \omega = 2\pi f = \frac{2\pi}{T} \qquad c = \lambda f = \frac{\omega}{k}. \tag{6.22}$$

For P-waves, the displacement u is in the direction of wave propagation, and c equals the P-wave velocity α. Similarly, for S-waves, u is perpendicular to the propagation direction, and c equals the S-wave velocity β. Equation (6.21) for the displacement (u) can then be written as

$$u = A \sin(kx - \omega t) = A \sin k(x - ct). \tag{6.23}$$

The velocity c is also called the *phase velocity*. It is the velocity with which a constant phase (e.g., the "peak" or "trough," or one of the zero displacements) is transmitted. This can be seen by equating the phase to a constant and then differentiating the expression with respect to time, as follows:

$$kx - \omega t = \text{constant}$$
$$k\frac{dx}{dt} - \omega = 0$$
$$\frac{dx}{dt} = \frac{\omega}{k} = c. \tag{6.24}$$

To demonstrate that the displacement given by Eq. (6.23) is a solution of the one-dimensional wave equation (Eq. 6.4), we must partially differentiate u in Eq. (6.23) twice with respect to time (t) and twice with respect to position (x):

$$\frac{\partial u}{\partial x} = Ak\cos(kx - \omega t)$$
$$\frac{\partial^2 u}{\partial x^2} = -Ak^2\sin(kx - \omega t) = -k^2 u$$
$$\frac{\partial u}{\partial t} = -A\omega\cos(kx - \omega t)$$
$$\frac{\partial^2 u}{\partial t^2} = -A\omega^2\sin(kx - \omega t) = -\omega^2 u$$

$$\frac{\partial^2 u}{\partial t^2} = \frac{\omega^2}{k^2}\frac{\partial^2 u}{\partial x^2} = c^2\frac{\partial^2 u}{\partial x^2}. \tag{6.25}$$

For a P-wave traveling along the x-axis the dilatation θ is given by an equation similar to Eq. (6.25), with substitution of the P-wave velocity (α) for the velocity c. Similarly, for an S-wave along the x-axis (Eq. 6.14) the horizontal displacement w is given by an equation like Eq. (6.25), with appropriate substitutions of w for u and the S-wave velocity (β) for the velocity c. The solutions of the three-dimensional compressional and shear wave equations (Eqs. 6.19 and 6.20, respectively) are more complicated than those given by Eq. (6.23) because waves may travel in any direction through a medium that is not homogeneous.

6.2.2.4 D'Alembert's Principle

Equation (6.23) describes the particle displacement during the passage of a wave that is traveling in the direction of the positive x-axis with velocity c. Because the velocity enters the wave equation as c^2, the one-dimensional wave equation is also satisfied by the displacement

$$u = B \sin k(x + ct), \tag{6.26}$$

which corresponds to a wave traveling with velocity c in the direction of the negative x-axis.

In fact, any function of $(x \pm ct)$ that has first and second derivatives is a solution of the one-dimensional wave equation. This is known as *D'Alembert's principle*. It can be simply demonstrated for some generic wave shape $W(x - ct) = W(\phi)$ as follows:

$$\frac{\partial W}{\partial x} = \frac{\partial W}{\partial \phi}\frac{\partial \phi}{\partial x} = \frac{\partial W}{\partial \phi} \qquad \frac{\partial W}{\partial t} = \frac{\partial W}{\partial \phi}\frac{\partial \phi}{\partial t} = -c\frac{\partial W}{\partial \phi}$$

$$\frac{\partial^2 W}{\partial x^2} = \frac{\partial}{\partial x}\frac{\partial W}{\partial x} = \frac{\partial \phi}{\partial x}\frac{\partial}{\partial \phi}\frac{\partial W}{\partial x} = \frac{\partial^2 W}{\partial \phi^2}$$

$$\frac{\partial^2 W}{\partial t^2} = \frac{\partial}{\partial t}\frac{\partial W}{\partial t} = \frac{\partial \phi}{\partial t}\frac{\partial}{\partial \phi}\frac{\partial W}{\partial t} = -c\frac{\partial}{\partial \phi}\left(-c\frac{\partial W}{\partial \phi}\right) = c^2\frac{\partial^2 W}{\partial \phi^2}$$

$$\frac{\partial^2 W}{\partial t^2} = c^2\frac{\partial^2 W}{\partial x^2}. \tag{6.27}$$

Because Eq. (6.27) is valid for positive and negative values of c, its general solution W represents the superposition of waves traveling in opposite directions along the x-axis, and is given by

$$W = f(x - ct) + g(x + ct). \tag{6.28}$$

Equation (6.28) generalizes the wave equation solutions in Eqs. (6.23) and (6.26) to the case where the wave shape is not a simple sine function. In nature, the wave shape depends on the characteristics of the source; for example, the evolution of the rupture during an earthquake.

6.2.2.5 The Eikonal Equation

Consider a wave traveling with constant velocity c along the axis x', the orientation of which is given by the direction cosines (l, m, n). If x' is measured from the center of the coordinate axes (x, y, z) we can substitute $x' = lx + my + nz$ for x in Eq. (6.28). If we consider for convenience only the wave traveling in the direction of positive x', we get as the general solution to the wave equation

$$W = f(lx + my + nz - ct). \tag{6.29}$$

The progress of a wave is described by successive positions of its wavefront, defined as the surface in which all particles at a given instant in time are moving with the same phase. For a particular value of t, a constant phase of the wave equation solution given by Eq. (6.29) requires that

$$lx + my + nz = \text{constant}. \tag{6.30}$$

Equation (6.30) represents a family of planes perpendicular to a line with direction cosines (l, m, n). We began this discussion by describing a wave moving with velocity c along the direction x', and now we see that this direction is normal to the plane wavefronts. This is the direction that we earlier defined as the ray path of the wave.

The wave equation is a second-order differential equation. However, the wave shape function *W* is also a solution of a first-order differential equation. This is seen by differentiating *W* with respect to *x, y, z,* and *t,* respectively, which gives

$$\frac{\partial W}{\partial x} = \frac{\partial W}{\partial \phi}\frac{\partial \phi}{\partial x} = l\frac{\partial W}{\partial \phi}$$
$$\frac{\partial W}{\partial y} = \frac{\partial W}{\partial \phi}\frac{\partial \phi}{\partial y} = m\frac{\partial W}{\partial \phi}$$
$$\frac{\partial W}{\partial z} = \frac{\partial W}{\partial \phi}\frac{\partial \phi}{\partial z} = n\frac{\partial W}{\partial \phi}.$$

(6.31)

The direction cosines (*l, m, n*) are related by $l^2 + m^2 + n^2 = 1$, and so – as can be verified by substitution – the expressions in Eq. (6.31) satisfy the equation

$$\left(\frac{\partial W}{\partial x}\right)^2 + \left(\frac{\partial W}{\partial y}\right)^2 + \left(\frac{\partial W}{\partial z}\right)^2 = \left(\frac{\partial W}{\partial \phi}\right)^2.$$

(6.32)

The position x' where the wave is observed can be written in terms of the velocity *c* and the travel time τ (*x, y, z*), required by the wave to propagate from the coordinate origin to position (*x, y, z*):

$$x' = lx + my + nz = c\tau(x, y, z).$$

(6.33)

By differentiating with respect to *x*, we see that $l = c(\partial t/\partial x)$,with similar expressions for *m* and *n*. Substituting in Eq. (6.31), we get

$$\frac{\partial W}{\partial x} = c\left(\frac{\partial \tau}{\partial x}\right)\left(\frac{\partial W}{\partial \phi}\right) \qquad \frac{\partial W}{\partial y} = c\left(\frac{\partial \tau}{\partial y}\right)\left(\frac{\partial W}{\partial \phi}\right)$$

$$\frac{\partial W}{\partial z} = c\left(\frac{\partial \tau}{\partial z}\right)\left(\frac{\partial W}{\partial \phi}\right). \tag{6.34}$$

Substituting these equations into Eq. (6.32) and simplifying, we finally find

$$\left(\frac{\partial \tau}{\partial x}\right)^2 + \left(\frac{\partial \tau}{\partial y}\right)^2 + \left(\frac{\partial \tau}{\partial z}\right)^2 = \frac{1}{c^2}. \tag{6.35}$$

Equation (6.35) is called the *eikonal equation*. The word eikonal derives from the Greek word for image, and in seismology it is synonymous with the distance between the source of a wave and the location where it is recorded. Though derived here for a homogeneous medium, it can be shown that the eikonal equation remains valid for media where velocity c varies slowly compared to the wavelength of the wave. The eikonal equation establishes that changes in travel time from one position to another (left-hand side of Eq. (6.35)) are large when the velocity between the two positions is low, and vice versa. The surfaces $\tau(x, y, z)$ = constant represent the wavefronts, which, in a heterogeneous medium, are not generally planar, even at large distances from the source.

6.2.2.6 The Energy in a Seismic Wave

It is important to distinguish between the velocity with which a seismic wave travels through a material and the speed with which the particles of the material vibrate during the passage of the wave. The vibrational speed (v_p) of a harmonic wave is obtained by differentiating Eq. (6.23) with respect to time, which yields

$$v_{\mathrm{p}} = \frac{\partial u}{\partial t} = -\omega A \cos(kx - \omega t). \tag{6.36}$$

The *intensity* or *energy density* of a wave is the energy per unit volume in the wavefront and consists of kinetic and potential energy. The kinetic part is given by

$$I = \frac{1}{2}\rho v_{\mathrm{p}}^2 = \frac{1}{2}\rho\omega^2 A^2 \cos^2(kx - \omega t). \tag{6.37}$$

The energy density averaged over a complete harmonic cycle consists of equal parts of kinetic and potential energy; it is given by

$$I_{\mathrm{av}} = \frac{1}{2}\rho\omega^2 A^2, \tag{6.38}$$

i.e., the mean intensity of the wave is proportional to the square of its amplitude and of its frequency.

6.2.2.7 Attenuation of Seismic Waves

The further a seismic signal travels from its source, the weaker it becomes. The decrease of amplitude with increasing distance from the source is referred to as *attenuation*. It is partly due to the geometry of propagation of seismic waves, and partly due to viscoelastic properties of the material through which they travel.

The most important attenuation is due to *geometric spreading*. Consider the seismic body waves generated by a seismic source at a point P on the surface of a uniform half-space (see Fig. 6.2). The energy E_{b} in the wavefront at distance r from its source is distributed over the surface of a hemisphere with area $2\pi r^2$. The *intensity* (or energy density, I_{b}) of the body waves is the energy per unit area of the wavefront, and at distance *r* is:

$$I_\mathrm{b}(r) = \frac{E_\mathrm{b}}{2\pi r^2}. \tag{6.39}$$

A surface wave is constricted to spread out laterally. The disturbance affects not only the free surface but extends downwards into the medium to a depth *d*, which we can consider to be constant for a given wave (Fig. 6.2). When the wavefront of a surface wave reaches a distance *r* from the source, the energy E_s is distributed over a circular cylindrical surface with area $2\pi rd$. At a distance *r* from its source the intensity of the surface wave is given by:

$$I_\mathrm{s}(r) = \frac{E_\mathrm{s}}{2\pi rd}. \tag{6.40}$$

When the total energy, E_b or E_s, is conserved, these equations show that the decrease in intensity of body waves is proportional to $1/r^2$ while the decrease of surface wave intensity is proportional to $1/r$. As shown in Eq. (6.38), the intensity of a harmonic wave is proportional to the square of its amplitude. The corresponding amplitude decay of body waves and surface waves are proportional to $1/r$ and $1/\sqrt{r}$, respectively. Thus, seismic body waves are attenuated more rapidly than surface waves with increasing distance from the source.

This phenomenon is clearly visible in Fig. 6.1. The amplitude of body waves (P- and S-waves) is significantly smaller than the amplitude of Rayleigh (L_R) and Love (L_Q) surface waves.

Another reason for attenuation is the *absorption* of energy due to imperfect elastic properties. If the medium does not deform perfectly elastically, part of the energy in the wave is converted into heat. This type of attenuation of the seismic wave is referred to as viscoelastic *damping*.

The damping of seismic waves is described by the *quality factor* (*Q*), defined as the fractional loss of energy per cycle:

$$\frac{2\pi}{Q} = -\frac{\Delta E}{E}. \tag{6.41}$$

In this expression ΔE is the energy lost in one cycle and E is the total elastic energy stored in the wave. If we consider the damping of a seismic wave as a function of the distance that it travels, a cycle is represented by the wavelength (λ) of the wave, i.e., ΔE equals the change $\mathrm{d}E$ per increment of distance $\mathrm{d}r$, multiplied by the wavelength λ. Equation (6.41) can be rewritten for this case as

$$\frac{2\pi}{Q} = -\frac{1}{E}\lambda\frac{\mathrm{d}E}{\mathrm{d}r}$$

$$\frac{\mathrm{d}E}{E} = -\frac{2\pi}{Q}\frac{\mathrm{d}r}{\lambda}. \tag{6.42}$$

Damping can be measured by its effect on the amplitude of a seismic wave. We have seen that the energy in a wave is proportional to the square of its amplitude A (Eq. 6.38). Thus we can write $\mathrm{d}E/E = 2\,\mathrm{d}A/A$ in Eq. (6.42), and on solving we get the damped amplitude of a seismic wave at distance r from its source:

$$A = A_0\exp\left(-\frac{\pi}{Q}\frac{r}{\lambda}\right) = A_0\exp\left(-\frac{r}{D}\right). \tag{6.43}$$

In this equation $D = Q\lambda/\pi$ is the distance within which the amplitude falls to 1/e (36.8%, or roughly one-third) of its original value. The inverse of this distance (D^{-1}) is called the *absorption coefficient*. A rock with a high Q transmits a seismic wave with relatively little energy loss by absorption, and the distance D is large. Generally, Q for P-waves is larger than for S-waves. In the Earth's mantle, for instance, Q is of the order of 10,000 for P-waves and around 100 for S-waves. Because Q is a measure of the deviation from perfect elasticity, it is also encountered in the theory of free oscillations of the Earth (Section 6.2.4), and has an effect on fluctuations of the Earth's free rotation, as in the damping of the Chandler wobble.

It follows from Eq. (6.43) that the absorption coefficient (D^{-1}) is inversely proportional to the wavelength λ. Thus, the attenuation of a seismic wave by absorption is dependent upon the frequency of the signal. High frequencies are attenuated more rapidly than are lower frequencies, for the same propagation distance r. As a result, the frequency spectrum of a seismic signal changes as it travels through the Earth. Although the original signal may be a sharp pulse (resulting from a shock or explosion), the preferential loss of high frequencies as it travels away from the source causes the signal to assume a smoother shape. This selective loss of high frequencies by absorption is analogous to removing high frequencies from a sound source by the use of a filter. The Earth acts as a low-pass filter to seismic signals.

6.2.3 Seismic Surface Waves

A disturbance at the free surface of a medium propagates away from its source partly as seismic surface waves. There are two categories of surface waves, known as *Rayleigh* and *Love* waves. Just as P- and S-waves, they are distinguished from each other by their direction of particle motion (polarization). Historically, Rayleigh waves are symbolized by L_R (for long-period Rayleigh) and Love waves by L_Q (for long-period "Querwelle," the German word for transverse wave).

6.2.3.1 Rayleigh Waves

In 1885 Lord Rayleigh described the propagation of a surface wave along the surface of a semi-infinite elastic half-space. The particles in the wavefront of the Rayleigh wave are polarized in the vertical plane. The resulting particle motion can be regarded as a combination of the P- and SV-vibrations. If the propagation direction is to the right of the viewer (Fig. 6.6), the particle motion describes a *retrograde ellipse* in the vertical plane with its major axis vertical and minor axis in the direction of wave propagation. If the medium is a Poisson body, (i.e., Poisson's ratio $\nu = 0.25$, see Section 5.1.3.1), which is a good approximation for most of the Earth, the theory of Rayleigh waves gives, without derivation, a propagation velocity (V_R) equal to $\sqrt{2-2/\sqrt{3}} \approx 0.92$ of the velocity (β) of S-waves (i.e., $V_R \approx 0.92\beta$).

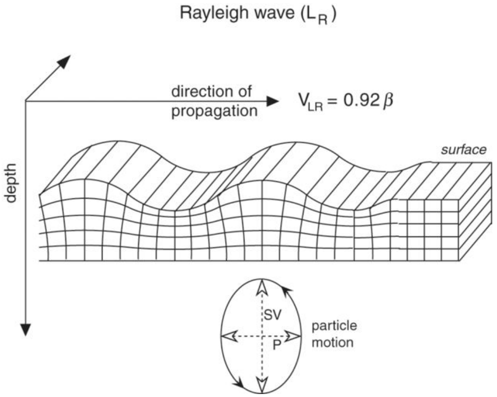


Fig. 6.6 The particle motion of a Rayleigh wave consists of a combination of P- and SV-vibrations in the vertical plane. The particles move in retrograde sense around an ellipse that has its major axis vertical and minor axis in the direction of wave propagation.

The particle displacement is not confined entirely to the surface of the medium. Particles below the surface are also affected by the passage of a Rayleigh wave; in a uniform half-space the amplitude of the particle displacement decreases exponentially with increasing depth. The penetration depth of the surface wave is defined as the depth where the amplitude is attenuated to e^{-1} of its value at the surface. For Rayleigh waves with wavelength λ the characteristic penetration depth is about 0.4λ.

6.2.3.2 Love Waves

In a homogeneous elastic half-space, Rayleigh waves are the only type of surface wave. However, A. E. H. Love showed in 1911 that if a horizontal layer lies between the surface and the half-space (Fig. 6.7a), successively reflected SH-waves can interfere constructively to form a surface wave with horizontal particle motion (Fig. 6.7b). For this to happen, the velocity (β_1) of S-waves in the near-surface layer must be lower than in the underlying half-space (β_2). The velocity of the Love waves (V_L) lies between the two values: $\beta_1 < V_L < \beta_2$.

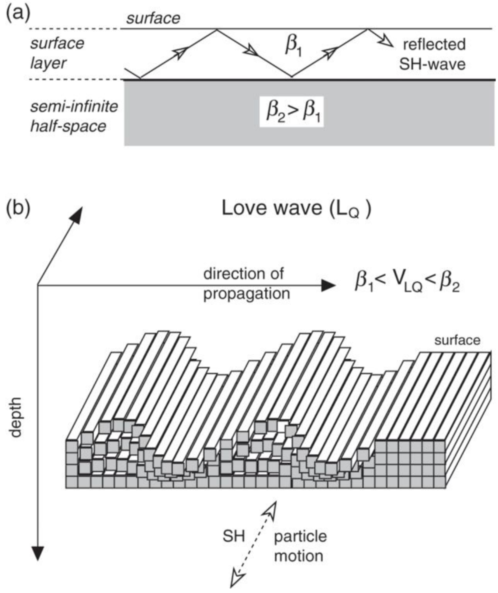

(a)
surface
surface layer
β1
reflected SH-wave
semi-infinite half-space
β2>β1
(b)
Love wave (LQ)
direction of propagation
β1< VLQ<β2
surface
depth
SH
particle motion

Fig. 6.7 (a) A Love wave results from constructive interference among SH-waves that are reflected at the top and bottom of the surface layer and become trapped in the layer. (b) The particle motion in a Love wave is horizontal and perpendicular to the direction of propagation. The amplitude of the wave decreases with depth below the free surface.

Theory shows that the velocity of Love waves with very short wavelengths is close to the slower velocity β_1 of the upper layer, while long wavelengths travel at a speed closer to the faster velocity β_2 of the lower medium. This dependence of velocity on wavelength is termed *dispersion.*

6.2.3.3 The Dispersion of Surface Waves

The dispersion of surface waves provides an important tool for determining the vertical velocity structure of the crust and upper mantle. Love waves are intrinsically dispersive even when the surface layer and underlying half-space are uniform. Rayleigh waves over a uniform half-space are non-dispersive. However, horizontal layers with different velocities are usually present or there is a vertical velocity gradient. Rayleigh waves with long wavelengths penetrate more deeply into the Earth than those with short wavelengths. The speed of Rayleigh waves is proportional to the shear-wave velocity ($V_R \approx 0.92\beta$), and in the crust and uppermost mantle β generally increases with depth. Thus, the deeper penetrating long wavelengths travel with faster seismic velocities than the short wavelengths. As a result, the Rayleigh waves are dispersive.

The packet of energy that propagates as a surface wave contains a spectrum of wavelengths. The energy in the wave propagates as the envelope of the wave packet (Fig. 6.8a), at a speed that is called the *group velocity* (U). The individual waves that make up the wave packet travel with *phase velocity* (c), as defined in Eq. (6.23). If the phase velocity is dependent on the wavelength, the group velocity is related to it by

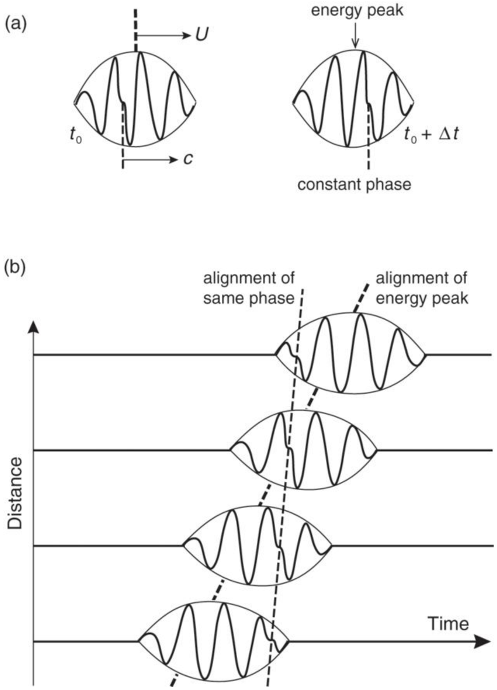


Fig. 6.8 (a) The surface wave energy propagates as the envelope of the wave packet with the group velocity U, while the individual wavelengths travel with the phase velocity c. (b) Change of shape of a wave packet due to normal dispersion as the faster-moving long wavelengths pass through the packet.

Source: W.M. Telford, L.P. Geldart and R.E. Sheriff, *Applied Geophysics*, 770 pp., Cambridge: Cambridge University Press, 1990.

$$U=\frac{\partial\omega}{\partial k}=\frac{\partial}{\partial k}(ck)$$

$$=c+k\frac{\partial c}{\partial k}=c-\lambda\frac{\partial c}{\partial\lambda}. \qquad (6.44)$$

The situation in which phase velocity increases with increasing wavelength (i.e., the longer wavelengths propagate faster than the short wavelengths) is called *normal dispersion*. In this case, because $\partial c/\partial\lambda$ is positive, the group velocity U is slower than the phase velocity c. The shape of the wave packet changes systematically as the faster-moving long wavelengths pass through the packet (Fig. 6.8b). As time elapses, an initially concentrated pulse becomes progressively stretched out into a long train of waves. Consequently, over a medium in which velocity increases with depth the long wavelengths arrive as the first part of the surface wave record at large distances from the seismic source.

The phenomenon of surface wave dispersion is clearly visible in our introductory seismogram example of Fig. 6.1. Love waves with a period of 10 s arrive later than Love waves at 30 s, which in turn are slower than 100 s Love waves.

6.2.4 Free Oscillations of the Earth

When a bell is struck with a hammer, it vibrates freely at a number of natural frequencies. The combination of free oscillations that are excited gives each bell its particular sonority. In an analogous way, the sudden release of energy in an earthquake sets the entire Earth into vibration, with natural frequencies of oscillation that are determined by the elastic properties and structure of the Earth's interior. Before discussing the Earth's free oscillations, it is worthwhile to review some concepts of vibrating systems that can be learned from the one-dimensional excitation of a vibrating string that is fixed at both ends.

Any complicated vibration of the string can be represented by the superposition of a number of simpler vibrations, called the *normal modes*. These arise when traveling waves reflected from the boundaries at the ends of the string interfere with each other to give a *standing wave*. Each normal mode corresponds to a standing wave with frequency and wavelength determined by the condition that the length of the string must always equal an integral number of half-wavelengths (Fig. 6.9). As well as the fixed ends, there are other points on the string that have zero displacement, called *nodes*. The first normal (*fundamental*) *mode* of vibration has no nodes. The second normal mode (*first overtone*) has one node; its wavelength and period are half those of the fundamental mode. The third normal mode (second overtone) has three times the frequency of the first mode, and so on. Modes with one or more node are also called *higher-order* modes.

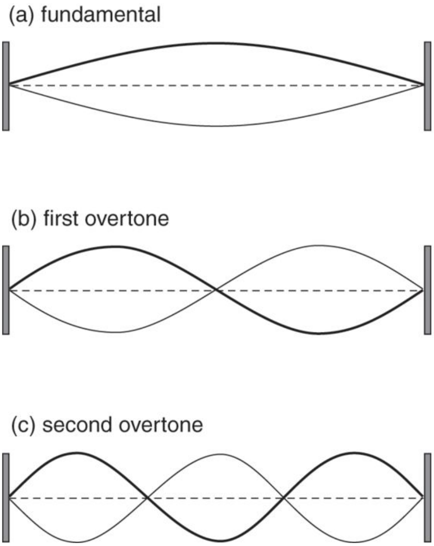


Fig. 6.9 Normal modes of vibration for a standing wave on a string fixed at both ends.

The concept of modes and nodes is also applicable to the vibration of a sphere, which can be resolved into a superposition of normal modes. The nodes of zero displacement become *nodal surfaces*, and instead of only one polarization direction there are three of them. In *spheroidal oscillations*, the volume of the Earth changes due to radial displacement. *Toroidal oscillations*, in contrast, are purely tangential, thus leaving the volume unperturbed.

6.2.4.1 Spheroidal Oscillations

The simplest kind of free oscillations are the *radial* oscillations, which are a special kind of spheroidal oscillations. In radial oscillations, the shape of the Earth remains “spherical” and all particles vibrate purely radially (Fig. 6.10a). In the fundamental mode of this type of oscillation the entire Earth expands and contracts in unison with a period of about 20.5 minutes. The second normal mode (first overtone) of radial oscillations has a single internal spherical nodal surface. While the inner sphere is contracting, the part outside the nodal surface is expanding, and vice versa. The nodal surfaces of higher modes are also spheres internal to the Earth and concentric with the outer surface.

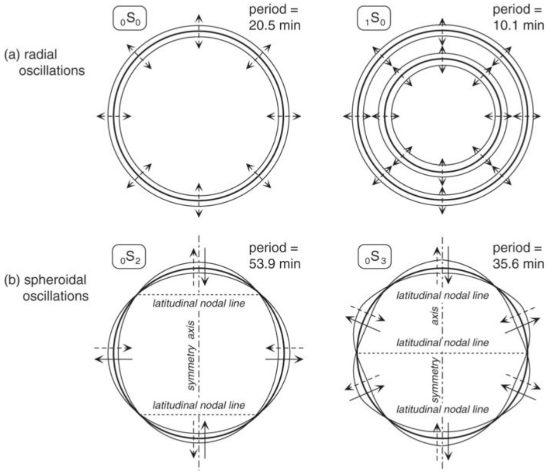

$_0S_0$
period = 20.5 min
$_1S_0$
period = 10.1 min
(a) radial oscillations
$_0S_2$
period = 53.9 min
$_0S_3$
period = 35.6 min
(b) spheroidal oscillations
latitudinal nodal line
symmetry axis
latitudinal nodal line
latitudinal nodal line
axis
latitudinal nodal line
symmetry
latitudinal nodal line

Fig. 6.10 (a) Modes of radial oscillation with their periods. In the fundamental mode $_0S_0$ (also called the balloon mode) the entire Earth expands and contracts in unison; higher modes, such as $_1S_0$, have internal spherical nodal surfaces concentric with the outer surface. (b) Modes of general spheroidal oscillation $_0S_2$ (“football mode”) and $_0S_3$ with their periods; the nodal lines are small circles perpendicular to the symmetry axis. Radial oscillations are a special kind of spheroidal oscillation, where displacement is purely radial.

A general spheroidal oscillation involves both radial and tangential displacements that can be described by *spherical harmonic functions* (see Box 3.4). These functions are referred to an axis through the Earth at the point of interest (e.g., an earthquake epicenter), and to a great circle which contains the axis. With respect to this reference frame they describe the latitudinal and longitudinal variations of the displacement of a surface from a sphere. They allow complete mathematical description and concise identification of each mode of oscillation with the aid of three indices. The longitudinal *order* m is the number of nodal lines on the sphere that are great circles, the *degree* l is determined from the $(l - m)$ latitudinal nodal lines, and the *overtone number* n describes the number of internal nodal surfaces. The notation ${}_nS_l^m$ denotes a spheroidal oscillation of degree l, longitudinal order m, and overtone number n. In practice, only oscillations with longitudinal order $m=0$ (rotationally symmetric about the reference axis) are observed, and this index is usually dropped. Also, the oscillation of degree $l = 1$ does not exist; it would have only a single equatorial nodal plane and the vibration would involve displacement of the center of gravity. The spheroidal oscillations ${}_0S_2$ and ${}_0S_3$ are shown in Fig. 6.10b. Spheroidal oscillations displace the Earth's surface and alter the internal density distribution.

The *radial oscillations* are spheroidal oscillation with $l=0$. The fundamental radial oscillation is also the fundamental spheroidal oscillation, denoted ${}_0S_0$; the next higher mode is called ${}_1S_0$ (Fig. 6.10a). Furthermore, in the context of the normal modes, l, m, n are often referred to as the *angular, azimuthal*, and *radial* orders.

6.2.4.2 Toroidal Oscillations

The third category of free oscillations is characterized by displacements that are purely tangential. The spherical shape and volume of the Earth are unaffected by a toroidal oscillation, which involves only longitudinal displacements about an axis (Fig. 6.11a). The amplitude of the longitudinal displacement varies with latitude. (Note that, as for spheroidal oscillations, "latitude" and "longitude" refer to the symmetry axis and are different from their geographic definitions). The toroidal modes have nodal planes that intersect the surface on circles of "latitude," on which the toroidal displacement is zero. Analogously to the spheroidal oscillations, the notation ${}_nT_l$ is used to describe toroidal modes. The mode ${}_0T_0$ has zero displacement and ${}_0T_1$ does not exist, because it describes a constant azimuthal twist of the entire Earth, which would change its angular momentum. The simplest toroidal mode is ${}_0T_2$, in which two hemispheres oscillate in opposite senses across a single nodal plane (Fig. 6.11a). Higher toroidal modes of degree l oscillate across $(l-1)$ nodal planes perpendicular to the symmetry axis.

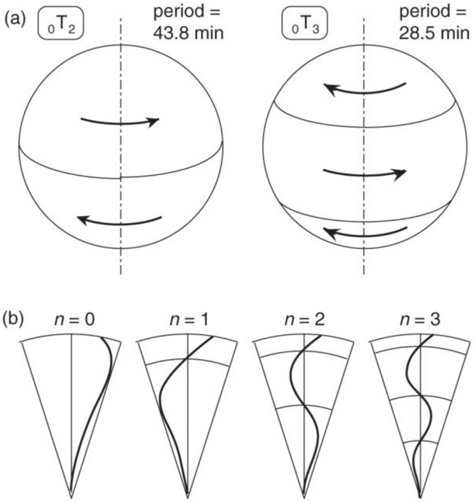


Fig. 6.11 (a) The modes of toroidal oscillation ${}_0T_2$ and ${}_0T_3$ with their periods; these modes involve oscillation in opposite senses across nodal planes normal to the symmetry axis. (b) In toroidal oscillations, displacements of internal spherical surfaces from their equilibrium positions vary with depth and are zero at internal nodal surfaces.

The amplitudes of toroidal oscillations change with depth. The displacements of internal spherical surfaces are zero at internal nodal surfaces (Fig. 6.11b). The nomenclature for toroidal oscillations also represents these internal nodal surfaces. Thus, ${}_1T_l$ denotes a general toroidal mode with one internal nodal surface; the internal sphere twists in an opposite sense to the outer spherical shell; ${}_2T_l$ has two internal nodal surfaces, etc. (Fig. 6.11b).

The twisting motion in toroidal oscillations does not alter the radial density distribution in the Earth. It depends on the shear strength of the Earth's interior. The Earth's fluid core cannot take part in these oscillations, and they are therefore restricted to the Earth's rigid mantle and crust.

6.2.4.3 Comparison with Surface Waves

The higher-order free oscillations of the Earth are related directly to the two types of surface wave.

1. In a Rayleigh wave the particle vibration is polarized in the vertical plane and has radial and tangential components (see Fig. 6.6). The higher-order *spheroidal* oscillations are equivalent to the standing wave patterns that arise from the interference of long-period Rayleigh waves traveling in opposite directions around the Earth.

2. In a Love wave the particle vibration is polarized horizontally. The *toroidal* oscillations may be regarded as the standing wave patterns due to the interference of oppositely traveling Love waves.

The similarity between surface waves and higher-order natural oscillations of the Earth is evident in the variations of displacement with depth. Like any vibration, a surface wave is made up of different modes. Theoretical analysis of surface waves shows that the amplitudes of different modes decay with depth in the Earth (Fig. 6.12) in an equivalent manner to the depth attenuation of natural oscillations (Fig. 6.11b).

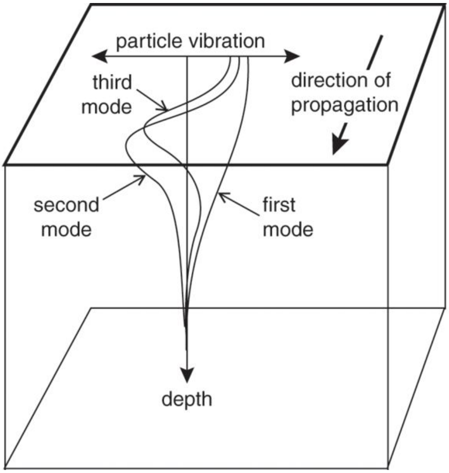


Fig. 6.12 The attenuation with depth of some low-order modes of Love waves.

Due to their low amplitudes and very long periods, free oscillations were calculated for spherically symmetric Earth models long before they could actually be observed. The great 1960 Chile earthquake (magnitude around 9.5, one of the largest ever recorded) marked a turning point in free oscillation research. The frequencies of more than 40 fundamental modes could be identified. Since then, thanks to steady improvements of recording instruments (seismometers), our knowledge of the Earth's free oscillations has improved substantially, and they can now be used to investigate the three-dimensional heterogeneous structure of the planet.

Figure 6.13 illustrates some of the free oscillations of the Earth set up by the 2004 Sumatra–Andaman Islands earthquake (magnitude 9.0). Vertical motions were recorded over 240 hours by the Canberra, Australia, station of the Geoscope global network of broadband seismic stations (see Section 6.5). The power spectrum in Fig. 6.13 is a plot of the energy associated with different frequencies of vibration. The splitting of some normal modes into two or more peaks is caused by the non-spherical shape and rotation of the Earth.

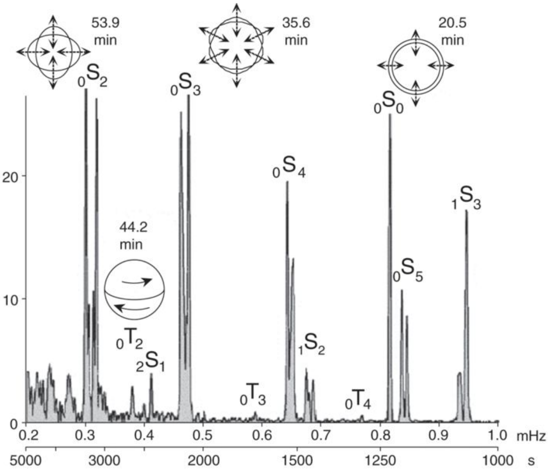


Fig. 6.13 Spectrum of natural oscillations of the Earth following the magnitude 9.0 Sumatra–Andaman earthquake of December 26, 2004.

Source: J. Park, T.A. Song et al., Earth's free oscillations excited by the 26 December 2004 Sumatra–Andaman earthquake, *Science*,

1139–1144, 2005.

The study of the free oscillations of the Earth set up by large earthquakes is an important branch of seismology, because the normal modes are strongly dependent on the Earth's internal structure. The periods of free oscillation are determined by the radial distributions of elastic properties and density in the Earth. Comparison of the observed periods of different modes with values computed for different models of the Earth's velocity and density structure provides an important check on the validity of the Earth model.

The oscillations are damped by the anelasticity of the Earth. The low-order oscillations with periods of up to 40–50 minutes persist much longer than the higher-order oscillations with short periods of only a few minutes. By studying the decay times of different modes, a profile of the quality factor (Q) within the Earth can be calculated that is compatible with results obtained for body waves.

6.2.5 Seismic Anisotropy

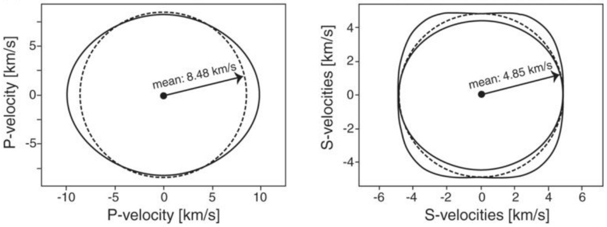
The strength of atomic bonds in most crystals that compose crustal and mantle rocks differs from one direction to another. As a consequence, the deformation of a crystal depends on the direction in which stress is applied. This directional dependence is referred to as *anisotropy*. Crystals that do not exhibit anisotropy are said to be *isotropic*. Anisotropy means that the elastic parameters, and therefore the velocity of seismic waves, are directionally dependent.

While isotropic materials can be completely described with only two elastic parameters, the shear and the bulk modulus, for example, anisotropic materials require up to 21 elastic parameters. Though the physics and mathematics of anisotropy are well understood, we limit ourselves to a phenomenological account.

When all mineral grains in a rock are randomly oriented, the anisotropy of individual grains averages out, and seismic wave propagation appears macroscopically isotropic. Coherent deformation of large volumes of rock, much larger than a seismic wavelength, can lead to a collective preferred orientation of many grains in a particular direction. This phenomenon, known as *lattice-preferred orientation* (LPO), occurs for example as a consequence of mantle flow, which tends to align the crystal lattice along the flow direction. The consequence is macroscopic anisotropy of the Earth's crust and mantle that manifests itself in the properties of seismic waves.

For body waves, anisotropy leads to a directional dependence of the propagation velocity. There are directions in which body waves travel faster than in others. Furthermore, the single S-wave that propagates through an isotropic medium splits into two separate S-waves with different velocities (Fig. 6.14). This *shear wave splitting* can be observed for S-waves recorded at the majority of seismic stations, and it may serve as a proxy for flow in the mantle.

(a) Olivine



(b) Muscovite

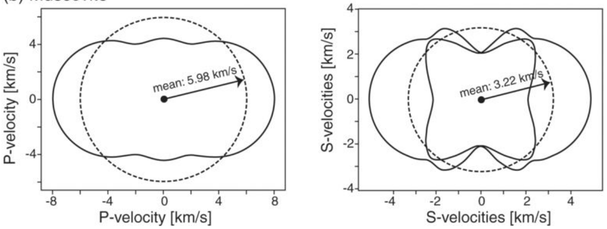


Fig. 6.14 Directional dependence of P-wave velocity (left) and the two S-wave velocities (right) in the anisotropic minerals olivine (a) and muscovite (b). The length of an arrow from the center to a point on the curve equals the velocity in the direction of the arrow. The dashed circle represents the velocity average over all directions, including the azimuthal directions, not shown in these two-dimensional plots.

Though shear wave splitting also occurs in surface waves, this effect is less easily observed because Love and Rayleigh waves often interfere, which complicates their individual analysis. However, it is well-established that arrival times of Love and Rayleigh waves often cannot be explained under the assumption of an isotropic Earth. Together with shear wave splitting, this observation, known as the *Love–Rayleigh discrepancy*, is the most robust indicator of anisotropy in the upper mantle.

6.3 Seismic Waves in the Heterogeneous Earth

6.3.1 Introduction

So far we have mostly considered waves traveling through a homogeneous medium. The Earth, however, is heterogeneous as a consequence of tectonic and magmatic processes in the crust, and convective motion in the mantle. Historically, two separate ways of describing wave propagation through heterogeneous media developed independently. One method, using Huygens' principle, describes the behavior of wavefronts; the other, using Fermat's principle, handles the geometry of ray paths. The eikonal equation (Section 6.2.2.5) establishes that these two methods are equivalent.

In the Earth the velocities of P- and S-waves are often proportional to each other. This follows from Eq. (6.7) and Eq. (6.15), which relate the body wave velocities to the Lamé constants λ and μ. For many rocks, Poisson's relation $\lambda = \mu$ approximately applies (see Box 5.1), and so

$$\frac{\alpha}{\beta} = \sqrt{\frac{\lambda + 2\mu}{\mu}} = \sqrt{3}. \tag{6.45}$$

The simplest form of heterogeneity is layering, and so most of our focus will be on the propagation of P-waves across the boundary between two layers with P-wave velocities α_1 and α_2. We can equally apply the analyses to S-waves, by substituting the appropriate shear wave velocities β_1 and β_2.

6.3.2 Huygens' Principle

The passage of a wave through a medium and across interfaces between adjacent media was explained by the seventeenth-century Dutch mathematician and physicist, Christiaan Huygens, who formulated a principle for the propagation of light as a wave, rather than as the stream of particles visualized by his great and influential contemporary, Sir Isaac Newton. Although derived for optics, Huygens' principle (1678) can be applied equally to any kind of wave phenomenon. The theory is based on simple geometrical constructions and permits the future position of a wavefront to be calculated if its present position is known. Huygens' principle can be stated as: "All points on a wavefront can be regarded as point sources for the production of new spherical waves; the new wavefront is the tangential surface (or envelope) of the secondary wavelets."

This principle can be illustrated simply for a plane wavefront, although it also applies to curved wavefronts (Fig. 6.3). Let the wavefront initially occupy the position AB (Fig. 6.15) and let the open circles represent individual particles of the material in the wavefront. The particles are agitated by the arrival of the wavefront and act as sources of secondary wavelets. Assuming a P-wave with propagation velocity α, the distance traveled by a wavelet after time t is αt. Each wavelet describes a small sphere around its source. If the original wavefront contained numerous closely spaced particles instead of a discrete number, the plane CD tangential to the small wavelets would represent the new position of the wavefront. It is also planar, and lies at a perpendicular distance αt from the original wavefront. In their turn the particles in the wavefront CD act as sources for new secondary wavelets, and the process is repeated. This principle can be used to derive the laws of reflection and refraction of seismic waves at an interface, and also to describe the process of *diffraction* by which a wave is deflected at a corner or at the edge of an object in its path.

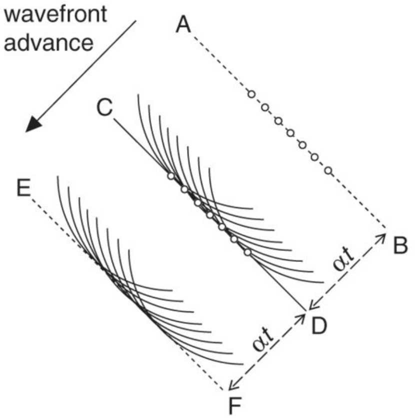


Fig. 6.15 Application of Huygens' principle to explain the advance of a plane wavefront. The wavefront at CD is the envelope of wavelets set up by particle vibrations when the wavefront was at the previous position AB. Similarly, the envelope of wavelets set up by vibrating particles in the wavefront CD forms the wavefront EF.

6.3.2.1 The Law of Reflection Using Huygens' Principle

Consider a plane P-wave traveling in a medium with seismic velocity α_1 when it encounters the boundary to another medium with velocity is α_2 (Fig. 6.16). At the boundary part the energy of the incident wave is transferred to the second medium, and the remainder is reflected back into the first medium. If the incident wavefront AC first makes contact with the interface at A, it agitates particles of the first medium at A and simultaneously the particles of the second medium in contact with the first medium at A. The vibrations of these particles set up secondary waves that travel away from A, back into the first medium as a reflected wave with velocity α_1 (and onward into the second medium as a refracted wave with velocity α_2).

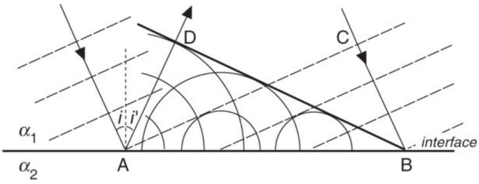


Fig. 6.16 The reflection of a plane P-wave at an interface between

two media with different seismic velocities. Dashed lines (e.g., AC) represent incident plane waves. Spherical wavelets are set up in the upper medium by vibrating particles in the segment AB of the interface. The reflected plane wave BD is the envelope of the wavelets.

By the time the incident wavefront reaches the interface at B all particles of the wavefront between A and B have been agitated. Applying Huygens’ principle, the wavefront of the reflected disturbance is the tangent plane to the secondary wavelet in the first medium. In Fig. 6.16 this is represented by the tangent BD from B to the circle centered at A, the first point of contact with the boundary. In the time t that elapses between the arrival of the plane wave at A and its arrival at B, the incident wavefront travels a distance CB and the secondary wavelet from A travels the equal distance AD. The triangles ABC and ABD are congruent. It follows that the reflected wavefront makes the same angle with the interface as the incident wave.

It is customary to describe the orientation of a plane by the direction of its normal. The angle between the normal to the interface and the normal to the incident wavefront is called the angle of incidence (i); the angle between the normal to the interface and the normal to the reflected wavefront is called the angle of reflection (i'). This application of Huygens’ principle to plane waves shows that the angle of reflection is equal to the angle of incidence ($i = i'$). This is known as the *law of reflection.*

6.3.2.2 The Law of Refraction Using Huygens’ Principle

The discussion of the interaction of the incident wave with a boundary can be extended to the transmission of the wave into the second medium (Fig. 6.17), where waves travel with velocity α_2. Let t be the time taken for the incident wavefront in the first medium to advance from C to B; then BC = $\alpha_1 t$. In this time all particles of the second medium between A and B have been agitated and now act as sources for new wavelets in the second medium. When the incident wave reaches B, the wavelet from A in the second medium has spread out to the point E, where AE = $\alpha_2 t$. The wavefront in the second medium is the tangent BE from B to the circle centered at A. The angle of incidence (i) is defined as before; the angle between the normal to the interface and the normal to the transmitted wavefront is called the angle of refraction (r). Comparison of the triangles ABC and ABE shows that BC = AB sin i, and AE = AB sin r. Consequently,

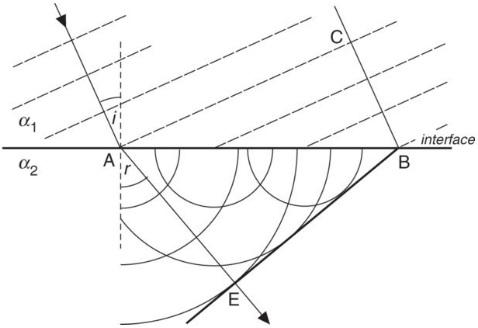


Fig. 6.17 The refraction of a plane P-wave at an interface between two media with different seismic velocities α_1 and α_2 ($> \alpha_1$). Dashed lines (e.g., AC) represent incident plane waves. Spherical wavelets are set up in the lower medium by vibrating particles in the segment AB of the interface. The refracted plane wave BE is the envelope of the wavelets. The angles of incidence (i) and refraction (r) are defined between the normal to the interface and the respective rays.

$$\frac{AB\sin i}{AB\sin r}=\frac{BC}{AE}=\frac{\alpha_1 t}{\alpha_2 t}, \tag{6.46}$$

$$\frac{\sin i}{\sin r}=\frac{\alpha_1}{\alpha_2}. \tag{6.47}$$

Equation (6.47) is called the *law of refraction*, also known as *Snell’s law*, in recognition of the Dutch mathematician Willebrod Snellius (or Snell).

6.3.2.3 Diffraction

When a plane or spherical seismic wave encounters a pointed obstacle or discontinuous surface, it experiences *diffraction*. This phenomenon allows the wave to bend around the obstacle, penetrating what otherwise would be a shadow zone for the wave. It is the diffraction of sound waves, for example, that allows us to hear the voices of people who are still invisible to us around a corner. Huygens’ principle also gives an explanation for diffraction, as illustrated by the following simple case.

Consider the normal incidence of a plane wave on a straight boundary that ends at a sharp corner B (Fig. 6.18). The incident wave is reflected along the entire length AB, with each particle of AB acting as a secondary source according to Huygen’s principle. Beyond the edge B the incident wavefronts cannot be reflected. The plane wavefront passes by the edge B, so that point C should lie in the shadow of AB. However, the corner B also acts as a source of secondary wavelets, part of which contribute to the reflected wavefront and part pass into the shadow zone. The intensity of the wave diffracted into the shadow zone is weaker than in the main wavefront, and it decreases progressively with increasing angle away from the direction of travel of the incident wavefront.

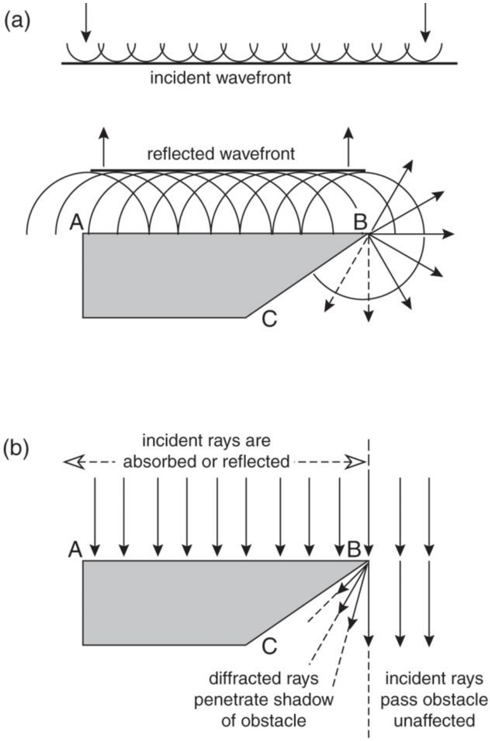

(a)
incident wavefront
reflected wavefront
A
B
C
(b)
incident rays are absorbed or reflected
A
B
C
diffracted rays penetrate shadow of obstacle
incident rays pass obstacle unaffected

Fig. 6.18 Explanation of diffraction at an edge with the aid of Huygens' principle. (a) The incident and reflected plane wavefronts are the envelopes to Huygens wavelets, which are able to carry the incident disturbance around a sharp corner. (b) The incident rays are absorbed, reflected or pass by the obstacle, but some rays related to the wavelets generated at the point of the obstruction are diffracted into its shadow.

6.3.3 Fermat’s Principle

The behavior of seismic ray paths at an interface is explained by another principle originally formulated in optics by the seventeenth-century French mathematician Pierre de Fermat. Fermat’s principle states that, of the many possible paths between two points A and B, the ray follows the path along which the travel time is a minimum. If d*s* is the element of distance along a ray path and α is the P-wave velocity over this short distance, then the geometry of the ray path is such that the travel time *t* between A and B is *minimal*, i.e.,

$$t = \int_{A}^{B} \frac{\mathrm{d}s}{\alpha} = \text{minimum.} \tag{6.48}$$

The velocity α may vary continuously with position. For arbitrary velocity variations the calculation of the ray path is intricate. In the case of a layered medium, Fermat’s principle provides us with an independent method for determining the laws of reflection and refraction. Of course, Fermat’s principle also holds for S-waves, in which case α should be replaced by β in Eq. (6.48).

6.3.3.1 The Law of Reflection Using Fermat’s Principle

Consider the reflection of a ray in a medium with constant P-wave velocity α_1 at the boundary to another medium (Fig. 6.19). Let A be a point on the incident ray at a vertical distance *h* from the boundary and let B be the corresponding point on the reflected ray. Let C and D be the nearest points on the boundary to A and B, respectively. Further, let *d* be the horizontal separation AB, and let O be the point of reflection on the interface at a horizontal distance *x* from C; then OD is equal to $(d - x)$ and we can write for the travel time *t* from A to B:

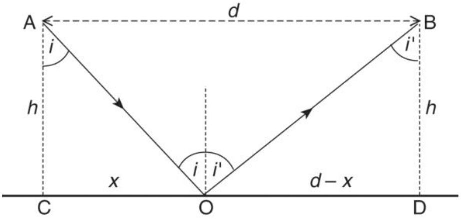


Fig. 6.19 Geometry of incident and reflected rays for derivation of the law of reflection with the aid of Fermat's principle.

$$t = \frac{AO}{\alpha_1} + \frac{OB}{\alpha_1} = \frac{1}{\alpha_1}\left[\sqrt{h^2 + x^2} + \sqrt{h^2 + (d - x)^2}\right]. \tag{6.49}$$

According to Fermat's principle, the travel time *t* must be a minimum. The only variable in Eq. (6.49) is *x*. To find the condition that gives the minimum travel time we differentiate *t* with respect to *x* and set the result equal to zero:

$$\frac{dt}{dx} = \frac{1}{\alpha}\left[\frac{x}{\sqrt{h^2 + x^2}} - \frac{(d - x)}{\sqrt{h^2 + (d - x)^2}}\right]. \tag{6.50}$$

By inspection of Fig. 6.19 the relationships of these expressions to the angle of incidence (*i*) and the angle of reflection (*i′*) are evident. The first expression inside the brackets is sin*i* and the second is sin*i′*. The condition for the minimum travel time is again $i' = i$; the angle of reflection equals the angle of incidence.

6.3.3.2 The Law of Refraction Using Fermat's Principle

We can use a similar approach to determine the law of refraction. This time we study the passage of the ray from a medium with velocity α_1

into a medium with higher velocity α_2 (Fig. 6.20). Let A again be a point on the incident ray at a vertical distance h from a point C on the interface. The ray traverses the boundary at O, a horizontal distance x from C. Let B now be a point on the ray in the second medium at a distance h from D, the closest point on the interface. The distance CD is d, so that again OD is equal to $(d - x)$. The travel time t that we have to minimize is given by

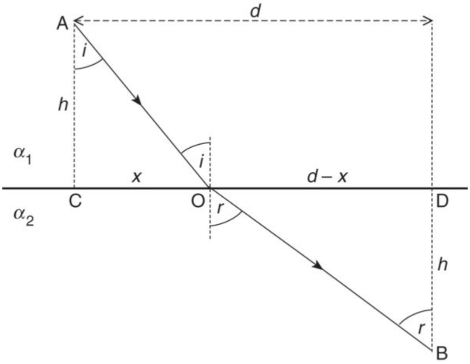


Fig. 6.20 Geometry of incident and refracted rays for derivation of the law of refraction with the aid of Fermat's principle.

$$t = \frac{AO}{\alpha_1} + \frac{OB}{\alpha_2} = \frac{\sqrt{h^2 + x^2}}{\alpha_1} + \frac{\sqrt{h^2 + (d - x)^2}}{\alpha_2}. \tag{6.51}$$

Differentiating Eq. (6.51) with respect to x and setting the result equal to zero gives us the condition for the minimal value of t:

$$\frac{x}{\alpha_1\sqrt{h^2 + x^2}} - \frac{(d - x)}{\alpha_2\sqrt{h^2 + (d - x)^2}} = 0. \tag{6.52}$$

By reference to Fig. 6.20 we can write this expression in terms of the sines of the angles of incidence (i) and refraction (r). This application of Fermat's principle to the ray paths gives again the law of refraction that we derived by applying Huygens' principle (Eq. 6.47). It can also be stated as

$$\frac{\sin i}{\alpha_1} = \frac{\sin r}{\alpha_2}. \tag{6.53}$$

In this example we have assumed that $\alpha_2 > \alpha_1$. As it passes from the medium with lower velocity into the medium with higher velocity the refracted ray is bent away from the normal to the boundary, giving an angle of refraction that is greater than the angle of incidence ($r > i$). Under the opposite conditions, if $\alpha_2 < \alpha_1$, the refracted ray is bent back toward the normal and the angle of refraction is less than the angle of incidence ($r < i$).

Fermat's principle as formulated in Eq. (6.48) is in fact a simplification of a more general version that asserts that the geometry of a ray path may also be such that the travel time is a maximum. In this situation any small change of the ray path connecting A and B actually decreases the travel time by a little bit. While such interesting cases occur for example in the presence of low-velocity anomalies in the Earth, they can be largely ignored in our present analysis of wave propagation in layered media.

6.3.4 Partitioning of Seismic Body Waves at a Boundary

The conditions that must be fulfilled at a boundary are that the normal and tangential components of stress, as well as the normal and tangential components of the displacements, must be continuous across the interface. If the normal (or tangential) stress were not continuous, the point of discontinuity would experience infinite acceleration. Similarly, if the normal displacements were not continuous, a gap would develop between the media or parts of both media would overlap to occupy the same space; discontinuous tangential displacements would result in relative motion between the media across the boundary. These anomalies are impossible if the boundary is a fixed surface that clearly separates the media.

As a result of the conditions of continuity, a P-wave incident on a boundary energizes the particles on each side of the boundary at the point of incidence, and sets up four waves. The energy of the incident P-wave is partitioned between P- and S-waves that are reflected from the boundary, and other P- and S-waves that are transmitted into the adjacent layer. The way in which this takes place may be understood by considering the particle motion that is induced at the interface.

The particle motion in the incident P-wave is parallel to the direction of propagation. At the interface, the vibration of particles of the lower layer can be resolved into a component perpendicular to the interface and a component parallel to it in the vertical plane containing the incident P-wave. In the second layer each of these motions can in turn be resolved into a component parallel to the direction of propagation (a refracted P-wave) and a component perpendicular to it in the vertical plane (a refracted SV-wave). Because of continuity at the interface, similar vibrations are induced in the upper layer, corresponding to a reflected P-wave and a reflected SV-wave, respectively.

Let the angles between the normal to the interface and the ray paths of the P- and S-waves in medium 1 be i_p and i_s, respectively, and the corresponding angles in medium 2 be r_p and r_s (Fig. 6.21). Applying Snell's law to both the reflected and refracted P- and S-waves gives

$$\frac{\sin i_p}{\alpha_1} = \frac{\sin i_s}{\beta_1} = \frac{\sin r_p}{\alpha_2} = \frac{\sin r_s}{\beta_2}. \tag{6.54}$$

By similar reasoning it is evident that an incident SV-wave also generates vibrations that have components normal and parallel to the interface, and will set up refracted and reflected P- and SV-waves. The situation is different for an incident SH-wave, which has no component of motion normal to the interface. In this case only refracted and reflected SH-waves are created.

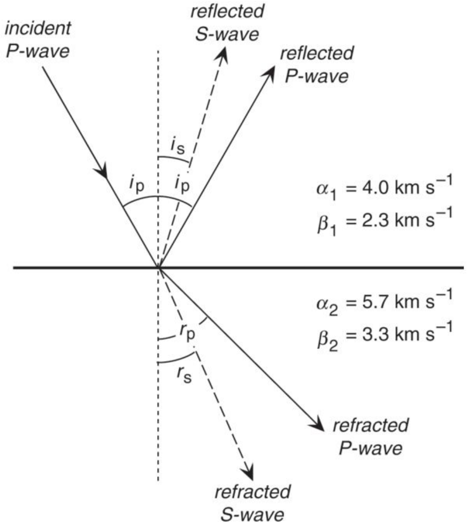

incident
P-wave
reflected
S-wave
reflected
P-wave
i_s
i_p
i_p
α_1 = 4.0 km s^{-1}
β_1 = 2.3 km s^{-1}
α_2 = 5.7 km s^{-1}
β_2 = 3.3 km s^{-1}
r_p
r_s
refracted
P-wave
refracted
S-wave

Fig. 6.21 The generation of reflected and refracted P- and S-waves from a P-wave incident on a plane interface. Dashed lines mark the rays of waves transmitted into the lower layer.

6.3.4.1 Subcritical and Supercritical Reflections, and Critical Refraction

Let O be a seismic source near the surface of a uniformly thick horizontal layer with P-wave velocity α_1 that lies on top of a layer with higher velocity α_2 (Fig. 6.22). Consider what happens to seismic rays that leave O and arrive at the boundary with all possible angles of incidence. The most simple ray is that which travels vertically to meet the boundary with zero angle of incidence at point N. This *normally incident ray* is partially reflected back along its track, and partially transmitted vertically into the next medium without change of direction. As the angle of incidence increases, the point of incidence moves from N toward C. The transmitted ray experiences a change of direction according to Snell's law of refraction, and the ray reflected to the surface is termed a *subcritical reflection*.

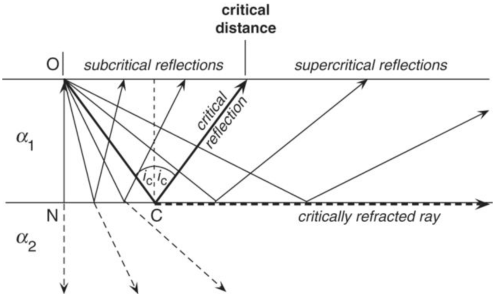


Fig. 6.22 The critical reflection defines two domains, corresponding to regions of subcritical and supercritical reflection, respectively.

The ray that is incident on the boundary at C is called the *critical ray* because it experiences *critical refraction*. It encounters the boundary with a *critical angle of incidence*. The corresponding refracted ray makes an angle of refraction of 90° with the normal to the boundary. As a result, it travels parallel to the boundary in the top of the lower layer with faster velocity α_2. The sine of the angle of refraction of the critical ray is unity, and we can calculate the critical angle, i_c, by applying Snell's law:

$$\sin i_c = \frac{\alpha_1}{\alpha_2}. \tag{6.55}$$

The critical ray is accompanied by a *critical reflection*. It reaches the surface at a *critical distance* (x_c) from the source at O. The reflections that arrive inside the critical distance are called *subcritical reflections*. At angles up to the critical angle refracted rays pass into the lower medium, but for rays incident at angles greater than the critical angle refraction is no longer possible. The seismic rays that are incident more obliquely than the critical angle are reflected almost completely. These reflections are termed *supercritical reflections*, or simply *wide-angle reflections*. They lose little energy to refraction, and are thus capable of traveling large distances from the source in the upper medium. Supercritical reflections are recorded with strong amplitudes on seismograms at distant stations.

6.3.5 Reflection of Seismic Waves in Layered Media

Using the law of reflection we can predict the travel times of waves propagating through a stratified medium. We first consider the simplest case of a single horizontal interface before analyzing more complicated scenarios with multiple or inclined layers. The resulting equations allow us to infer the depths and velocities of subsurface layers from observations of reflection travel times at the surface.

6.3.5.1 Reflection at a Horizontal Interface

The simplest case of seismic reflection is the two-dimensional reflection at a horizontal boundary (Fig. 6.23). Let the reflector be at depth d below the shot-point S. The ray that strikes the boundary at R is reflected to the surface and recorded at point G, so that the angles of incidence and reflection are equal. Let G be at a horizontal distance x from the shot-point. If the P-wave velocity is α, the first signal received at G is from the direct wave that travels directly along SG. Its travel time is given by $t_\mathrm{d} = x/\alpha$. It is important to keep in mind that the direct wave is not a surface wave but a body wave that travels parallel to and just below the surface of the top layer. The travel time t of the reflected ray SRG is $(\mathrm{SR} + \mathrm{RG})/\alpha$. However, SR and RG are equal and therefore

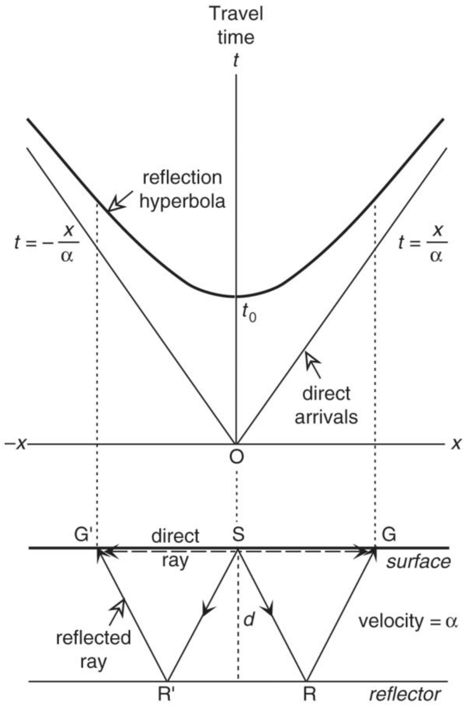

Travel time t
reflection hyperbola
$t = -\frac{x}{\alpha}$
$t = \frac{x}{\alpha}$
t_0
direct arrivals
−x
O
x
G'
direct ray
S
G
surface
d
reflected ray
velocity = α
R'
R
reflector

Fig. 6.23 The travel time versus distance curve for reflections from a horizontal boundary is a hyperbola. The vertical reflection time t_0 is the intercept of the hyperbola with the travel time axis.

$$t = \frac{2}{\alpha}\sqrt{d^2 + \frac{x^2}{4}}, \tag{6.56}$$

$$t = \frac{2d}{\alpha}\sqrt{1 + \frac{x^2}{4d^2}} = t_0\sqrt{1 + \frac{x^2}{4d^2}}. \tag{6.57}$$

At $x = 0$ the travel time corresponds to the vertical echo from the reflector; this echo-time is given by $t_0 = 2d/\alpha$. The quantity under the square root in Eq. (6.57) determines the curvature of the t–x curve and is called the *normal moveout* factor. It arises because the ray reaching a receiver at a horizontal distance x from the shot-point has not traveled vertically between it and the reflector. Squaring both sides of Eq. (6.57) and rearranging terms gives

$$\frac{t^2}{t_0^2} - \frac{x^2}{4d^2} = 1. \tag{6.58}$$

This is the equation of a hyperbola (Fig. 6.23) that is symmetrical about the vertical time axis, which it intersects at t_0. For large distances from the shot-point ($x \gg 2d$) the travel time of the reflected ray approaches the travel time of the direct ray and the hyperbola is asymptotic to the two lines $t = \pm x/\alpha$.

A principle goal of seismic reflection profiling is usually to find the vertical distance (d) to a reflecting interface. This can be determined from t_0, the two-way reflection travel time recorded at the shot-point, once the velocity α is known. One way of determining the velocity is by comparing t_0 with the travel time t_x to a receiver (seismometer), at distance x. In reflection seismology the receivers are often laid out close to the shot-point and the assumption is made that the receiver distance is much less than the depth of the reflector ($x \ll d$). Equation (6.57) becomes

$$t_x = t_0\sqrt{1+\left(\frac{x}{2d}\right)^2} = t_0\left[1+\frac{1}{2}\left(\frac{x}{2d}\right)^2+\right] \\ \approx t_0\left[1+\frac{1}{2}\left(\frac{x}{\alpha t_0}\right)^2\right]. \tag{6.59}$$

The difference between the travel time t_x and the shot-point travel time t_0 is the *normal moveout*, $\Delta t_n = t_x - t_0$. By rearranging Eq. (6.59), we get

$$\Delta t_n = \frac{x^2}{2\alpha^2 t_0}. \tag{6.60}$$

The echo time t_0 and the normal moveout time Δt_n are found from the reflection data. The distance x of the receiver from the shot-point is known and therefore the layer velocity α can be determined. The depth d of the reflector can then be found by using the formula for the echo time.

An alternative way of interpreting reflection arrival times becomes evident when Eq. (6.58) is rearranged in the form

$$t^2 = t_0^2 + \frac{x^2}{\alpha^2}. \tag{6.61}$$

A plot of t^2 against x^2 is a straight line that has slope $1/\alpha^2$. Its intercept with the t^2-axis gives the square of the echo time, t_0, from which the depth d to the reflector can be found once the velocity α is known. The record at each receiver will contain reflections from several reflectors. For the first reflector the velocity determined by the t^2–x^2 method is the true *interval velocity* of the uppermost layer, α_1, which, in conjunction with t_{01}, the first echo time, gives the thickness d_1 of the top layer. However, the ray reflected from the second interface has traveled through the first layer with interval velocity α_1 and the second layer with interval velocity α_2. The velocity interpreted in the t^2–x^2 method for this reflection, and for reflections from all deeper interfaces, is an *average velocity*. If the incident and reflected rays travel nearly vertically, the average velocity $\alpha_{a,n}$ for the reflection from the nth reflector is given by

$$\alpha_{\mathrm{a,n}} = \frac{d_1 + d_2 + d_3 + \cdots \quad d_n}{t_1 + t_2 + t_3 + \cdots \quad t_n} = \frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} t_i}, \tag{6.62}$$

where d_i is the thickness and t_i the interval travel time for the *i*th layer.

The t^2–x^2 method is a simple way of estimating layer thicknesses and average velocities for a multilayered Earth (Fig. 6.24). The slope of the second straight line gives $\alpha_{\mathrm{a},2}$, which is used with the appropriate echo time t_{02} to find the combined depth D_2 to the second interface, given by $D_2 = d_1 + d_2 = (\alpha_{\mathrm{a},2})(t_{02})$; d_1 is known, and so d_2 can be calculated. The two-way travel time in the second layer is $(t_{02} - t_{01})$ and thus the interval velocity α_2 can be found. In this way the thicknesses and interval velocities of deeper layers can be determined successively.

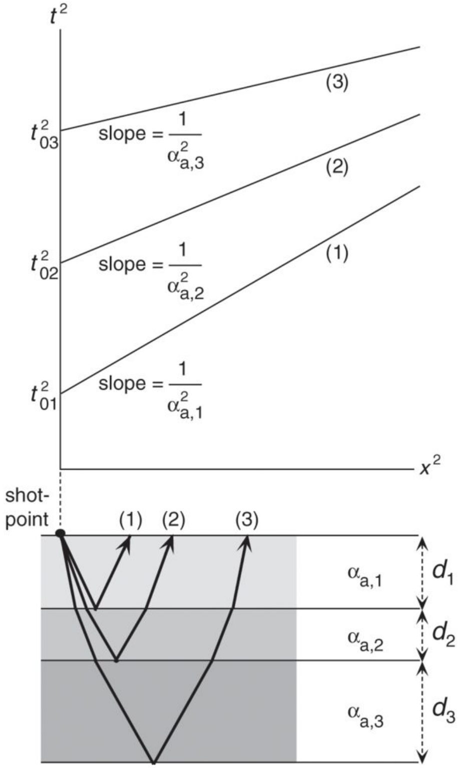

t^2
t^2_{03}
t^2_{02}
t^2_{01}
slope = $\frac{1}{\alpha^2_{a,3}}$
slope = $\frac{1}{\alpha^2_{a,2}}$
slope = $\frac{1}{\alpha^2_{a,1}}$
(3)
(2)
(1)
x^2
shot-point
(1) (2) (3)
$\alpha_{a,1}$
$\alpha_{a,2}$
$\alpha_{a,3}$
d_1
d_2
d_3

Fig. 6.24 Illustration of a "t^2–x^2 plot" for near-vertical reflections from three horizontal reflectors; $\alpha_{a,1}$ is the true velocity α_1 of layer 1, but $\alpha_{a,2}$ and $\alpha_{a,3}$ are "average" velocities that depend on the true velocities and the layer thicknesses.

6.3.5.2 Reflection at an Inclined Interface

When the reflecting interface is inclined at an angle θ to the horizontal, as in Fig. 6.25, the shortest distance d between the shot-points and the reflector is the perpendicular distance to the inclined plane. The paths of reflected rays on the down-dip side of the shot-point are longer than those on the up-dip side; this has corresponding effects on the travel times. The rays obey the laws of reflection optics and appear to return to the surface from point S′, which is the image point of the shot-point with respect to the reflector. The travel time t through the layer with velocity α is readily found with the aid of the image point. For example, for the ray SRG, recorded by a receiver on the surface at G, we get

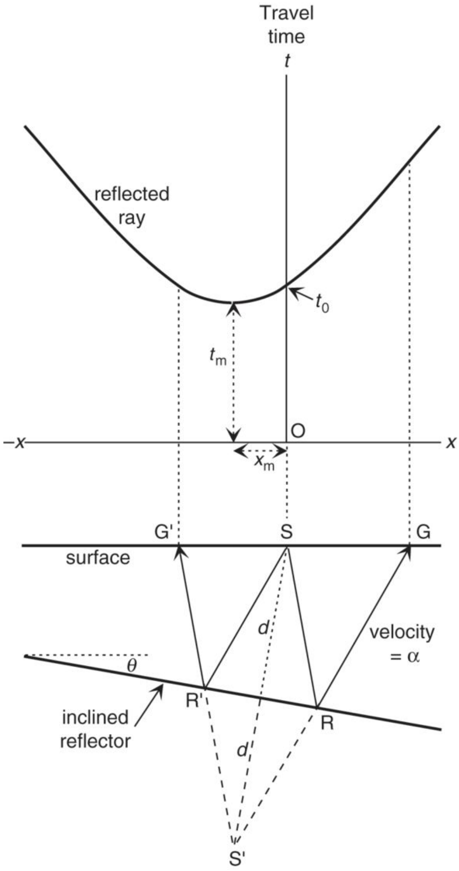

Travel
time
t
reflected
ray
t_0
t_m
$-x$
O
x
x_m
G'
S
G
surface
d
velocity
$= \alpha$
θ
R'
R
inclined
reflector
d
S'

Fig. 6.25 The travel time versus distance curve for an inclined reflector is also a hyperbola with vertical axis (cf. Fig. 6.23), but the minimum travel time (t_m) is measured at distance x_m from the shot-point.

$$t = \frac{\mathrm{SR} + \mathrm{RG}}{\alpha} = \frac{\mathrm{S'R} + \mathrm{RG}}{\alpha} = \frac{\mathrm{S'G}}{\alpha}. \tag{6.63}$$

The image point S′ is as far behind the reflector as the shot-point is in front: S′S = $2d$. In triangle S′SG the side SG equals the receiver distance x and the obtuse angle S′SG equals ($90^\circ + \theta$). If we apply the law of cosines to the triangle S′SG we can solve for S′G and substitute the answer in Eq. (6.63). This gives the travel time of the reflection from the inclined boundary:

$$t = \frac{1}{\alpha}\sqrt{x^2 + 4xd\ \sin\theta + 4d^2}. \tag{6.64}$$

Equation (6.64) is the equation of a hyperbola whose axis of symmetry is vertical, parallel to the *t*-axis. For a flat reflector the hyperbola was symmetric about the *t*-axis (Fig. 6.23) and the minimum travel time (echo time) corresponded to the vertical reflection below the shot-point ($x = 0$). For an inclined reflector the minimum travel time t_m is no longer the perpendicular path to the reflector, which would give the travel time t_0 in Fig. 6.25. Although the perpendicular path is the shortest distance from the shot-point to *reflector*, it is not the shortest path of a reflected ray between the shot-point and a receiver. The shortest travel time is recorded by the receiver at a horizontal distance x_m on the up-dip side of the shot-point. The coordinates (x_m, t_m) of the minimum point of the travel time hyperbola are

$$x_m = -2d\ \sin\theta$$

$$t_m = \frac{2d\ \cos\theta}{\alpha}. \tag{6.65}$$

In practice it is not known a priori whether a reflector is horizontal or inclined. If reflection records are not corrected for the effect of layer-dip, an error results in plotting the positions of dipping beds. The shot-point travel time t_0 gives the direct distance to a reflector, but the path along which the echo has traveled is not known. Consider the geometry of the inclined boundary in Fig. 6.26. First-arrival reflections recorded for shot-points P, Q, and R come from the true reflection points A, B, and C. If the computed reflector depths are plotted directly below the shot-points at A′, B′, and C′, the dipping boundary will appear to lie at a shallower depth than its true position, and the apparent dip of the reflector will be less steep than the true dip. This leads to a distorted picture of the subsurface structure. For example, an anticline appears broader and less steep-sided than it is. Similarly, if the limbs of a syncline dip steeply enough, the first arrivals from the dipping limbs can conceal the true structure (Fig. 6.27).

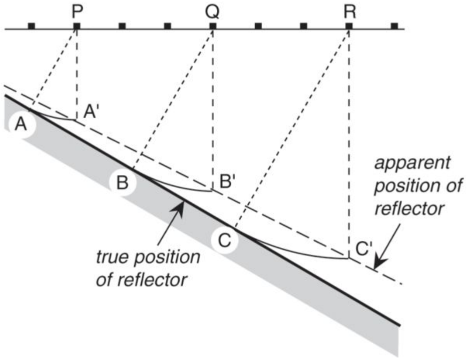

P
Q
R
A'
A
B
B'
C
C'
apparent position of reflector
true position of reflector

Fig. 6.26 When the reflector is inclined and depths are plotted vertically under geophone positions, the true reflecting points A, B, and C are mapped at A′, B′, and C′, falsifying the position of the reflector.

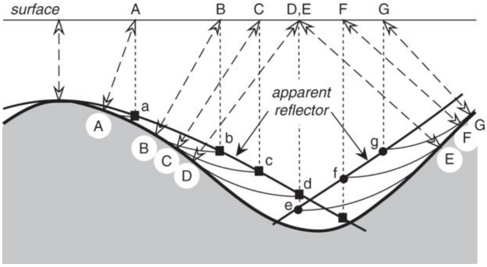

surface
A
B
C
D,E
F
G
apparent reflector
a
b
c
d
e
f
g
A
B
C
D
E
F
G

Fig. 6.27 Paths of reflected rays over an anticline and syncline, showing the false apparent depth to the reflecting surface. True reflection points A–G are wrongly mapped at locations a–g beneath the corresponding shot-points.

This happens when the radius of curvature of the bottom of the syncline is less than the subsurface depth of its axis. Over the axis of the syncline, rays reflected from the dipping flanks may be the first to reach the receiver positioned at the source location. The bottom of the syncline is seen as an upwardly convex reflection between two cusps (Fig. 6.28).

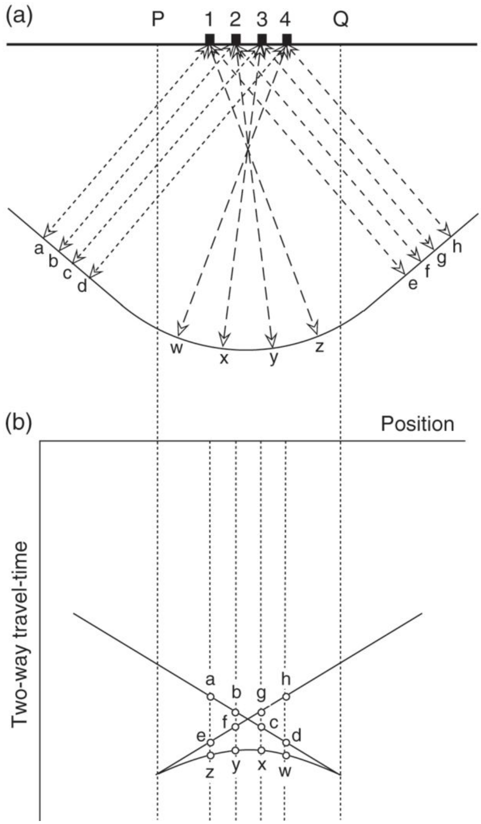


Fig. 6.28 (a) Paths of rays reflected from both flanks and the trough of a tightly curved syncline. (b) Appearance of the corresponding reflection record; the letters on the cusped feature refer to the reflection points in (a).

Source: H.A. Slack et al., The geomagnetic elements, *Geophysics* 32, 877–892, 1967.

Reflection seismic records must be corrected for non-vertical reflections. The correction process is called *migration*. It is an essential part of a reflection seismic study. When the reflection events on seismic cross-sections are plotted vertically below control points on the surface (e.g., as the two-way vertical travel-time to a reflector below the shot-point), the section is said to be *unmigrated*. As discussed above, an unmigrated section misrepresents the depth and dip of inclined reflectors. A *migrated* section is one which has been corrected for non-vertical reflections. It gives an improved image of the positions of subsurface reflectors.

The process of migration requires some prior knowledge of the seismic velocity distribution, which in an unexplored or tectonically complicated region is often inadequately known. Several techniques can be used, but to treat them adequately is beyond the scope of this text.

6.3.5.3 Reflection and Transmission Coefficients

The partitioning of energy between refractions and reflections at different angles of incidence on a boundary is rather complex. For example, an incident P-wave may be partially reflected and partially refracted, or it may be totally reflected, depending how steeply the incident ray encounters the boundary. The fraction of the incident P-wave energy that is partitioned between reflected and refracted P- and S-waves depends strongly on the angle of incidence (Fig. 6.29). In the case of oblique incidence at less than the critical angle, the amplitudes of the different waves are functions of the wave velocities and the angles of incidence, reflection, and refraction. The relative amounts of energy in the refracted and reflected P- and S-waves do not change much for angles of incidence up to about 15°. Beyond the critical angle, the refracted P-wave ceases, so that the incident energy is partly reflected as a P-wave and partly converted to refracted and reflected S-waves.

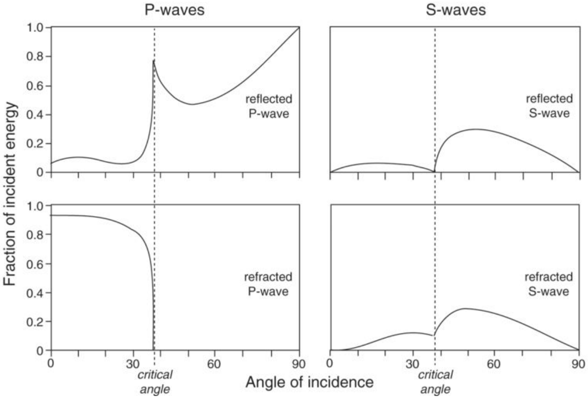


Fig. 6.29 Partitioning of the energy of an incident P-wave between refracted and reflected P- and S-waves (after Dobrin, 1976, and Richards, 1961).

In practice, reflection seismology is carried out at comparatively small angles of incidence. At *normal incidence* on an interface, a P-wave excites no tangential stresses or displacements, and no shear waves are induced. The partitioning of energy between the reflected and refracted P-waves then becomes much simpler. It depends on a property of each medium known as its *acoustic impedance, Z*, which is defined as the product of the density ρ of the medium and its P-wave velocity α; thus $Z = \rho\alpha$. The solution of the equations for the amplitudes A_1 and A_2 of the reflected and refracted P-waves, respectively, in terms of the amplitude A_0 of the incident wave are given by:

$$\mathrm{RC} = \frac{A_1}{A_0} = \frac{Z_2 - Z_1}{Z_2 + Z_1} = \frac{\rho_2\alpha_2 - \rho_1\alpha_1}{\rho_2\alpha_2 + \rho_1\alpha_1}$$

$$\mathrm{TC} = \frac{A_2}{A_0} = \frac{2Z_1}{Z_2 + Z_1} = \frac{2\rho_1\alpha_1}{\rho_2\alpha_2 + \rho_1\alpha_1}. \tag{6.66}$$

The amplitude ratios RC and TC are called the reflection coefficient and the transmission coefficient, respectively. As shown earlier (see Eq. 6.38), the energy of a wave is proportional to the square of its amplitude. The fraction E_r of the incident energy that is reflected is given by the square of RC; the fraction E_c that is transmitted is equal to $(1 - E_r)$.

When the incident wave is reflected at the surface of a medium with higher seismic impedance ($Z_2 > Z_1$), the reflection coefficient RC is positive. This means that the reflected wave is in phase with the incident wave. However, if the wave is incident on a medium with lower seismic impedance ($Z_2 < Z_1$), the reflection coefficient will be negative. This implies that the reflected wave is 180° out of phase with the incident wave. The fraction of energy reflected from an interface is equal to RC^2, and therefore does not depend on whether the incidence is from the medium of higher or lower seismic impedance.

Given some knowledge or preliminary guess of the Earth's structure, the concepts of reflection and transmission at layer boundaries can be used to compute *synthetic seismograms*. The comparison of actually observed and synthetic seismograms is useful for separating real reflections from noise (Section 6.3.5.4) and for improving models of the Earth. Though the idea of computing synthetic seismograms is simple, the construction can be laborious. A vertically incident wave on the first boundary is partitioned into reflected and transmitted components, with amplitudes corresponding to the seismic impedances above and below the interface. The transmitted wave is further subdivided at the next deeper interface into other reflected and transmitted components, and this is repeated at each subsequent boundary. Each wave is followed as it is reflected and refracted at subsurface interfaces until it eventually returns to the surface. The theoretical record then consists of the superposition of the numerous waves. An alternative to the construction of synthetic seismograms by tracking reflections and transmissions is numerical wave propagation, introduced in Section 6.3.7.

6.3.5.4 Noise, Surface Waves, and Multiple Reflections

In addition to single reflections from layer boundaries, seismic recordings also contain noise, surface waves, and multiple reflections (waves that have been reflected more than once, e.g., by bouncing back and forth between interfaces). These additional signals may obscure single reflection signals used to infer subsurface properties.

Seismic noise can be categorized into *instrumental* and *ambient noise* (Section 6.4). Instrumental noise is caused by the recording instrument itself, and may result, for example, from the thermal motion of atoms. Instrumental noise is often *incoherent*; that is, it is nearly random and different at all receivers. Ambient noise, in contrast, consists of actual waves propagating through the Earth's interior. It may be excited by the interaction of ocean waves with the coastline, atmospheric pressure variations that act on the surface, or human and industrial activity. Ambient noise is usually *coherent* because it propagates as a well-defined wave between receivers. The ratio of the energy in the signal to that in the noise, called the *signal-to-noise ratio*, is a measure of the quality of a seismic record. The higher the signal-to-noise ratio the better the record; records that have a ratio less than unity are unlikely to be usable. Noise may be reduced by arranging receivers in groups or *arrays* and averaging (or *stacking*) the individual recordings to produce a single record. When n receivers form a group, this practice enhances the signal-to-noise ratio approximately by the factor $\sqrt{n}$.

Surface waves, historically called "*ground roll*" in exploration geophysics, can to some extent be removed from the recordings by choosing a specific receiver configuration. For example, if a group of receivers at equal distances cover a complete wavelength of the Rayleigh wave (Fig. 6.30a), the signals of individual geophones are effectively averaged to a low value. This procedure is only partially effective because surface waves in a highly complex near-surface layer may propagate in multiple directions. Another method of suppressing surface waves is frequency filtering. The frequency of the surface waves is often lower than that of the reflected P-waves, which allows attenuation of surface waves by high-pass filtering (removal of lower frequencies).

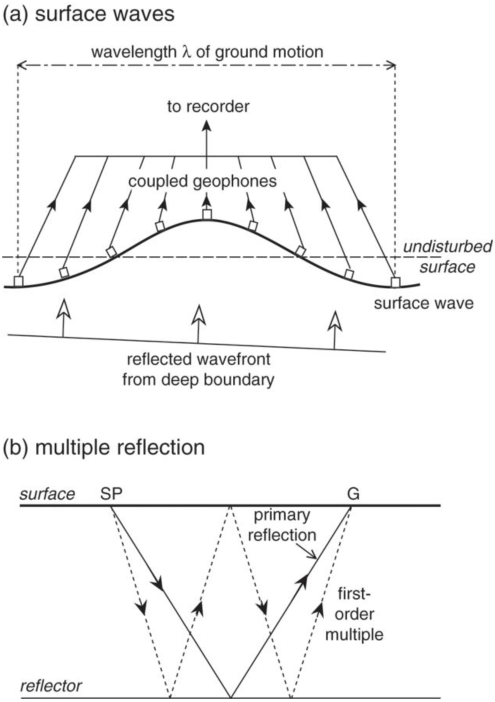

(a) surface waves
wavelength λ of ground motion
to recorder
coupled geophones
undisturbed surface
surface wave
reflected wavefront from deep boundary
(b) multiple reflection
surface
SP
G
primary reflection
first-order multiple
reflector

Fig. 6.30 Examples of (a) surface wave and (b) multiple reflections between a reflector and the free surface, which may obscure single reflection recordings.

Multiple reflections can originate in several ways, the most important of which is repeated reflection at the surface (Fig. 6.30b). The reflection coefficient at the surface is high (in principle, $RC \approx -1$), and multiple reflections can occur between the surface and a reflecting interface at depth. The travel time for a first-order multiple at near-vertical incidence is double that of the primary signal. A copy of the reflector is observed on the seismogram at twice the real depth (or travel time) when the multiple is incorrectly interpreted as a single reflection. Higher-order multiples produce additional apparent reflectors.

6.3.5.5 Reflection Seismic Section

After migration, the cleaned records from all receivers may be plotted side by side to form a sort of cross-section of the subsurface structure. Strong reflectors can be followed across the section, and where they are interrupted, faults can be deduced. The top part of Fig. 6.31 shows the results of a 113 km long, nearly north–south crustal reflection profile across western Lake Superior at the northern end of the North American Mid-Continent Rift System. This is an aborted Precambrian (age ~1000 Ma) rift that has prominent gravity and magnetic expressions. The profile was carried out with closely spaced sources (explosions). The records were subsequently stacked and migrated. The lower part of the figure shows the interpreted subsurface structures above a rather steeply dipping Moho, which under this profile increases in depth from about 32 km in the north to about 50 km in the south. The section shows the presence of some inferred major faults, such as the southward dipping Douglas fault and the northward dipping Keweenawan fault, which appears to truncate other steeply dipping faults.

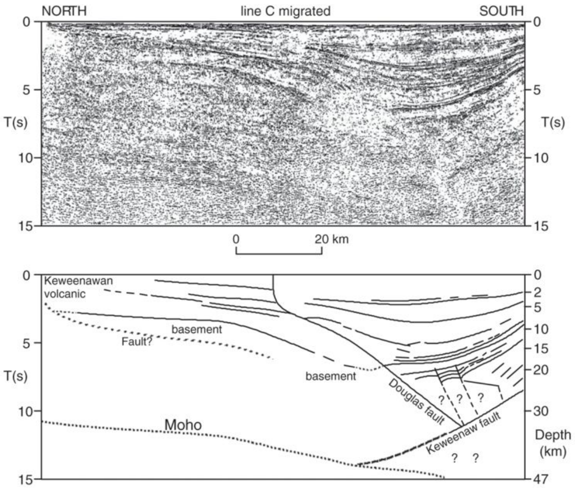


Fig. 6.31 Results of deep reflection profiling across the North American Mid-Continent Rift System under western Lake Superior. The top part of the figure shows the migrated reflection record, the bottom part the interpreted crustal structure (courtesy of A. G. Green).

6.3.6 Refraction of Seismic Waves in Layered Media

Seismic refraction can be understood by applying Huygens' principle to the critical refraction at the interface between two layers. The seismic wave travels immediately below the interface with the higher velocity of the lower medium. It is called a *head wave* (or *Mintrop wave*, after the German seismologist who patented its use in seismic exploration in 1919). The upper and lower media are in contact at the interface and so the upper medium is forced to move in phase with the lower medium. The vibration excited at the boundary by the passage of the head wave acts as a moving source of secondary waves in the upper layer. The secondary waves interfere constructively (in the same way as a reflected wave is formed) to build plane wavefronts; the ray paths return to the surface at the critical angle within the region of supercritical reflections (Fig. 6.32). The doubly refracted waves are especially important for the information they reveal about the layered structure of the deep interior of the Earth.

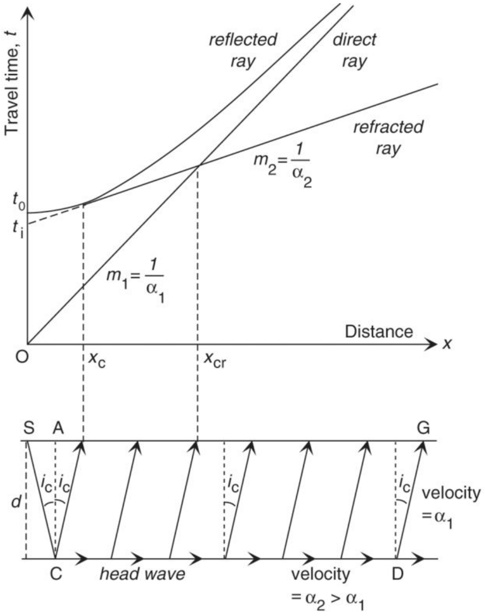

Travel time, t
reflected ray
direct ray
refracted ray
$m_2 = \frac{1}{\alpha_2}$
t_0
t_i
$m_1 = \frac{1}{\alpha_1}$
Distance
x
O
x_c
x_{cr}
S
A
G
i_c
i_c
d
i_c
i_c
velocity $= \alpha_1$
C
head wave
velocity $= \alpha_2 > \alpha_1$
D

Fig. 6.32 Travel time versus distance curves for the direct ray and the reflected and refracted rays at a horizontal interface between

two layers with seismic velocities α_1 and α_2 ($\alpha_1 > \alpha_2$).

6.3.6.1 Refraction at a Horizontal Interface

Refraction along a flat interface between two horizontal layers is illustrated in Fig. 6.32. Let the depth to the interface be d and the seismic P-wave velocities of the upper and lower layers be α_1 and α_2, respectively ($\alpha_1 < \alpha_2$). (As before, these arguments equally apply to S-waves when α is replaced by β.) The *direct ray* from the source at S is recorded by a receiver G at distance x on the surface after time x/α_1. The travel time curve for the direct ray is a straight line through the origin with slope $m_1 = 1/\alpha_1$. The hyperbolic t–x curve for the *reflected ray* intersects the time axis at the two-way vertical reflection ("echo") time t_0. At great distances from the shot-point the reflection hyperbola is asymptotic to the straight line for the direct ray.

The *doubly refracted ray* travels along the path SC with the velocity α_1 of the upper layer, impinges with critical angle i_c on the interface at C, passes along the segment CD with velocity α_2 of the lower layer, and returns to the surface along DG with velocity α_1. The segments SC and DG are equal, CD = x – 2 SA and the travel time for the path SCDG can be written

$$t = \frac{2\mathrm{SC}}{\alpha_1} + \frac{\mathrm{CD}}{\alpha_2}, \tag{6.67}$$

i.e.,

$$t = \frac{2d}{\alpha_1 \cos i_c} + \frac{(x - 2d\tan i_c)}{\alpha_2}. \tag{6.68}$$

Rearranging terms and using Snell's law, $\sin i_c = \alpha_1/\alpha_2$, we get for the travel time of the doubly refracted ray:

$$t = \frac{x}{\alpha_2} + \frac{2d}{\alpha_1}\cos i_c. \tag{6.69}$$

The equation represents a straight line with slope $m_2 = 1/\alpha_2$. The doubly refracted rays are only recorded at distances greater than the critical distance x_c. The first arrival recorded at x_c can be regarded as both a doubly refracted ray and a reflection; the travel time line for the head wave is tangential to the reflection hyperbola at x_c. By backward extrapolation the refraction t–x curve is found to intersect the time axis at the *intercept time* t_i, given by

$$t_i = \frac{2d}{\alpha_1}\cos i_c = 2d\frac{\sqrt{\alpha_2^2 - \alpha_1^2}}{\alpha_1\alpha_2}. \tag{6.70}$$

Close to the source the direct ray is the first to be recorded. However, the doubly refracted ray travels part of its path at the faster velocity of the lower layer, so that it eventually overtakes the direct ray and becomes the first arrival. The straight lines for the direct and doubly refracted rays cross each other at this distance, which is accordingly called the *crossover distance,* x_{cr}. It is computed by equating the travel times for the direct and refracted rays:

$$\frac{x_{cr}}{\alpha_1} = \frac{x_{cr}}{\alpha_2} + 2d\frac{\sqrt{\alpha_2^2 - \alpha_1^2}}{\alpha_1\alpha_2}, \tag{6.71}$$

$$x_{cr} = 2d\sqrt{\frac{\alpha_2 + \alpha_1}{\alpha_2 - \alpha_1}}. \tag{6.72}$$

Using the travel times of refracted waves, the velocities of subsurface layers can be inferred directly from the reciprocal slopes of the straight lines corresponding to the direct and doubly refracted rays. Once these velocities have been determined, it is possible to compute the depth *d* to the interface by using either the intercept time t_i or the crossover distance x_{cr}, which can be read directly from the *t*–*x* plot:

$$d = \frac{1}{2} t_i \frac{\alpha_1 \alpha_2}{\sqrt{\alpha_2^2 - \alpha_1^2}}, \tag{6.73}$$

$$d = \frac{1}{2} x_{cr} \sqrt{\frac{\alpha_2 - \alpha_1}{\alpha_2 + \alpha_1}}. \tag{6.74}$$

6.3.6.2 Refraction at an Inclined Interface

In practice, the refracting interface is often not horizontal. The assumption of flat layers then leads to errors in the velocity and depth estimates. When the refractor is suspected to have a dip, the velocities of the beds and the dip of the interface can be obtained by shooting a second complementary profile in the opposite direction. Suppose a refractor dips at an angle θ as in Fig. 6.33. Sources A and B are located at the ends of a line of receivers that covers AB. The ray ACDB from the source A reaches the interface at the critical angle i_c at C, propagates as a head wave with velocity α_2 along the dipping interface, and the ray emerging at D eventually reaches a receiver at the end of the profile at B. In the reversed experiment the ray from the source at B to a geophone at A traverses the same path in the opposite direction. However, the *t*–*x* curves are different. Let d_A and d_B be the perpendicular distances from the sources A and B to the interface at P and Q, respectively. For the down-dip experiment at A the travel-time to distance x is given by

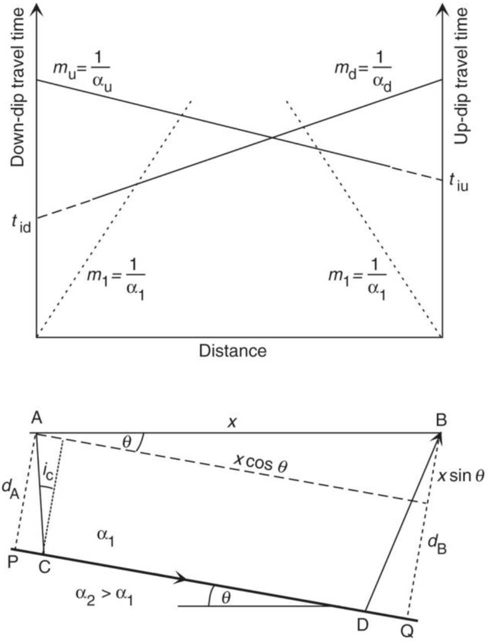

Down-dip travel time
Up-dip travel time
$m_u = \frac{1}{\alpha_u}$
$m_d = \frac{1}{\alpha_d}$
t_{iu}
t_{id}
$m_1 = \frac{1}{\alpha_1}$
$m_1 = \frac{1}{\alpha_1}$
Distance
A
x
B
θ
x cos θ
x sin θ
d_A
i_c
α_1
d_B
P
C
$\alpha_2 > \alpha_1$
θ
D
Q

Fig. 6.33 Travel time versus distance curves of direct and refracted rays for up-dip and down-dip profiles when the refracting boundary dips at angle θ.

$$t_d = \frac{\mathrm{AC} + \mathrm{DB}}{\alpha_1} + \frac{\mathrm{CD}}{\alpha_2}. \tag{6.75}$$

The geometry of Fig. 6.33 gives the following trigonometric relationships:

$$\begin{aligned} \mathrm{AC} &= \frac{d_A}{\cos i_c} & \mathrm{DB} &= \frac{d_B}{\cos i_c} \\ \mathrm{PC} &= d_A \tan i_c & \mathrm{DQ} &= d_B \tan i_c \\ \mathrm{CD} &= x\cos\theta - (\mathrm{PC} + \mathrm{DQ}); \end{aligned} \tag{6.76}$$

we have

$$t_d = \frac{(d_A + d_B)}{\alpha_1 \cos i_c} + \frac{x \cos\theta - (d_A + d_B)\tan i_c}{\alpha_2} = \frac{x \cos\theta}{\alpha_2} + \frac{(d_A + d_B)}{\alpha_1} \cos i_c. \tag{6.77}$$

Equation (6.77) can be simplified by noting that

$$d_B = d_A + x \sin\theta \qquad \text{and} \qquad \frac{1}{\alpha_2} = \frac{\sin i_c}{\alpha_1}. \tag{6.78}$$

After substitution and gathering terms, the down-dip travel time is given by

$$t_d = \frac{x \sin i_c \cos\theta}{\alpha_1} + \frac{x \cos i_c \sin\theta}{\alpha_1} + \frac{2d_A \cos i_c}{\alpha_1}, = \frac{x}{\alpha_1} \sin(i_c + \theta) + t_{id} \tag{6.79}$$

where t_{id} is the intercept time for the down-dip experiment:

$$t_{id} = \frac{2d_A}{\alpha_1} \cos i_c. \tag{6.80}$$

The analysis for the up-dip direction is analogous and gives

$$t_u = \frac{x}{\alpha_1} \sin(i_c - \theta) + t_{iu}, \tag{6.81}$$

where t_{iu} is the intercept time for the up-dip experiment:

$$t_{\mathrm{iu}} = \frac{2d_{\mathrm{B}}}{\alpha_1} \cos i_{\mathrm{c}}. \tag{6.82}$$

If the upper layer is homogeneous, the segments for the direct ray will have equal slopes, the reciprocals of which give the velocity α_1 of the upper layer. The segments of the curves corresponding to the doubly refracted ray are different for up-dip and down-dip experiments. The total travel times in either direction along ACDB must be equal, but the *t–x* curves have different intercept times. As these are proportional to the perpendicular distances to the refractor below the sources, the up-dip intercept time t_{iu} is larger than the down-dip intercept time t_{id}. This means that the slope of the up-dip refraction in Fig. 6.33 is flatter than the down-dip slope. If we interpret the reciprocal of the slope as the velocity of the lower medium, we get two *apparent velocities,* α_{d} and α_{u}, given by

$$\frac{1}{\alpha_{\mathrm{d}}} = \frac{1}{\alpha_1} \sin(i_{\mathrm{c}} + \theta) \qquad \frac{1}{\alpha_{\mathrm{u}}} = \frac{1}{\alpha_1} \sin(i_{\mathrm{c}} - \theta). \tag{6.83}$$

Once the actual velocity α_1 and the apparent velocities α_{d} and α_{u} have been determined from the *t–x* curves, the dip of the interface θ and the critical angle i_{c} (and from it the velocity α_2 of the lower layer) can be computed:

$$\theta = \frac{1}{2}\left[\sin^{-1}\left(\frac{\alpha_1}{\alpha_{\mathrm{d}}}\right) - \sin^{-1}\left(\frac{\alpha_1}{\alpha_{\mathrm{u}}}\right)\right], \tag{6.84}$$

$$i_{\mathrm{c}} = \frac{1}{2}\left[\sin^{-1}\left(\frac{\alpha_1}{\alpha_{\mathrm{d}}}\right) + \sin^{-1}\left(\frac{\alpha_1}{\alpha_{\mathrm{u}}}\right)\right]. \tag{6.85}$$

If the reciprocal apparent velocities (Eq. 6.83) are added, a simple approximation for the velocity of the lower layer is obtained:

$$\begin{aligned}\frac{1}{\alpha_{\mathrm{d}}}+\frac{1}{\alpha_{\mathrm{u}}}&=\frac{1}{\alpha_1}[\sin(i_{\mathrm{c}}+\theta)+\sin(i_c-\theta)]\\&=\frac{2}{\alpha_1}\sin i_{\mathrm{c}}\cos\theta\\&=\frac{2}{\alpha_2}\cos\theta.\end{aligned} \tag{6.86}$$

If the refractor dip is small, cos $\theta \approx 1$ (for example, if $\theta < 15°$, cos $\theta > 0.96$) and an approximate formula for the velocity of the second layer is

$$\frac{1}{\alpha_2}\approx\frac{1}{2}\left(\frac{1}{\alpha_d}+\frac{1}{\alpha_u}\right). \tag{6.87}$$

6.3.6.3 Refraction with Continuous Change of Velocity with Depth

Imagine the Earth to have a multilayered structure with numerous thin horizontal layers, each characterized by a constant seismic velocity, which increases progressively with increasing depth (Fig. 6.34). A seismic ray that leaves the surface with angle i_1 will be refracted at each interface until it is finally refracted critically. The ray that finally returns to the surface will have an emergence angle equal to i_1. Snell's law applies to each successive refraction (e.g., at the top surface of the nth layer, which has velocity α_n):

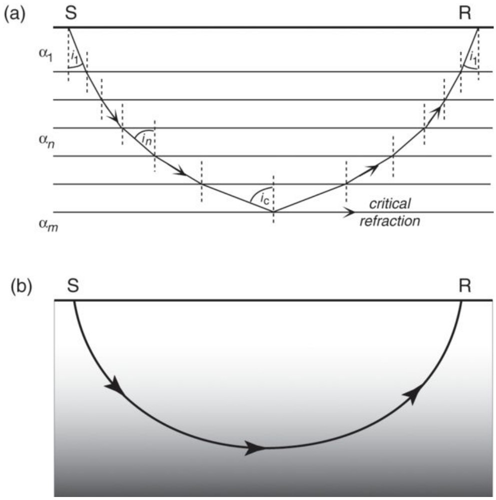


Fig. 6.34 (a) The path of a seismic wave through a horizontally layered medium, in which the seismic velocity is constant in each layer and increases with increasing depth, becomes ever flatter until critical refraction is reached; the return path of each emerging ray mirrors the incident path. (b) When the velocity increases continuously with depth, the ray is a smooth curve that is concave upward.

$$\frac{\sin i_1}{\alpha_1} = \frac{\sin i_2}{\alpha_2} = \cdots \quad \frac{\sin i_n}{\alpha_n} = \text{constant} = p. \tag{6.88}$$

The constant p is called the *ray parameter*. It is characteristic for a particular ray with emergence angle i_1 and velocity α_1 in the surface layer. If α_m is the velocity of the deepest layer, along whose surface the ray is eventually critically refracted ($\sin i_m = 1$), then the value of p must be equal to $1/\alpha_m$.

As the number of layers increases and the thickness of each layer decreases, the situation is approached in which the velocity increases constantly with increasing depth. Each ray then has a smoothly curved path. If the vertical increase of velocity is linear with depth, the curved rays are circular arcs.

In the above we have assumed that the refracting interfaces are horizontal, ignoring that the Earth has a nearly spherical shape. The passage of seismic body waves through a layered spherical Earth can be treated to a first approximation in the same way. We can represent the vertical (radial) velocity structure by subdividing the Earth into concentric shells, each with a faster body-wave velocity than the shell above it (Fig. 6.35). Snell's law of refraction applies to the interface between each pair of shells. For example, at point A we can write

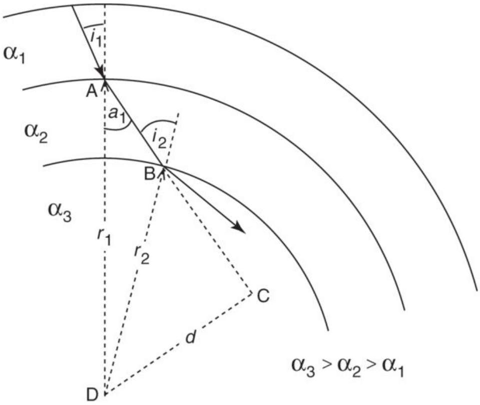


Fig. 6.35 Refraction of a seismic ray in a spherically layered Earth, in which the seismic velocity is constant in each layer and the layer velocity increases with depth.

$$\frac{\sin i_1}{\alpha_1} = \frac{\sin a_1}{\alpha_2}. \tag{6.89}$$

Multiplying both sides by r_1 gives

$$\frac{r \sin i_1}{\alpha_1} = \frac{r \sin a_1}{\alpha_2}. \tag{6.90}$$

In triangles ACD and BCD, respectively, we have

$$d = r_1 \sin a_1 = r_2 \sin a_2. \tag{6.91}$$

Combining Eqs. (6.89), (6.90), and (6.91) gives the result

$$\frac{r_1 \sin i_1}{\alpha_1} = \frac{r_2 \sin i_2}{\alpha_2} = \cdots \quad \frac{r_n \sin i_n}{\alpha_n} = \text{constant} = p_s. \tag{6.92}$$

The constant p_s is called the *spherical ray parameter*. Here, the seismic ray is a straight line within each spherical layer with constant velocity. If velocity increases continuously with depth, the seismic ray is refracted continuously and its shape is curved concavely upward. It reaches its deepest point when $\sin i = 1$, at radius r_0 where the velocity is α_0; these parameters are related by the *Benndorf relationship*:

$$\frac{r \sin i}{\alpha} = \frac{r_0}{\alpha_0} = p_s. \tag{6.93}$$

Determination of the spherical ray parameter is the key to determining the variation of seismic velocity inside the Earth. Access to the Earth's interior is provided by analysis of the travel times of seismic waves that have traversed the various internal regions and emerge at the surface, where they are recorded. We will see in Section 7.4.3.1 that the travel time (t) of a seismic ray to a known epicentral distance (Δ) can be mathematically inverted to give the velocity α_0 at the deepest point of the path.

6.3.7 Numerical Solutions of the Wave Equation

In the previous sections we mostly studied wave propagation through media that are either homogeneous or stratified. D'Alembert's principle states that solutions to the wave equation for a constant velocity are wave packets traveling to the left and to the right (Eq. 6.28). For stratified media we did not find such a simple equation for the wavefield, but we could still compute travel time curves. Explicit equations for the wavefield or for travel times, such as the ones we found so far, are called *analytical solutions*.

Though simple models of the subsurface with analytical solutions are convenient and important for our intuitive understanding of wave propagation, the interior of the Earth is more complex. In addition to lateral heterogeneities that often result from tectonic or magmatic processes, internal boundaries (e.g., between crust and mantle) are often irregular. Additional complications arise due to the topography of the Earth. Consequently, analytical solutions for these more complicated scenarios often do not exist, and we have to resort to *numerical approximations* or *numerical solutions* of the wave equation.

6.3.7.1 The Finite-Difference Method

Numerical solutions are based on the approximation of derivatives that appear in a differential equation. For this, a continuously defined function, such as u in the wave equations (Eq. 6.4 and 6.25), is replaced by a discrete set of values $u(x_1)$, $u(x_2)$, …, $u(x_k)$, sampled at some positions x_1, x_2, …, x_k, that are separated by some interval Δx. This replacement, illustrated in Fig. 6.36, is known as *discretization*. The choice of interval is important; if it is too narrow the amount of data is large, but if it is too wide, the continuous function may be poorly represented.

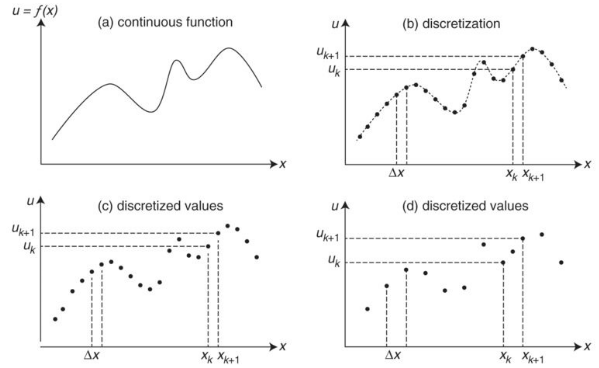


Fig. 6.36 (a) A continuous function that is to be converted to a set of discrete values. (b) Determination of discrete values of the function with a suitable choice of sampling interval. (c) The set of discrete values that portray the important features of the continuous function. (d) Loss of some features of the original function if the discretization interval is too wide.

The differential equation we want to discretize is the wave equation (Eq. 6.25). We write this as follows, to emphasize the dependence on the two variables, *x* and *t*:

$$\frac{\partial^2 u(x,t)}{\partial t^2} = c^2(x)\frac{\partial^2 u(x,t)}{\partial x^2}. \tag{6.94}$$

To allow for a variable velocity, *c* in Eq. (6.94) is now a function of location *x*. As before, it either represents P-wave velocity *α* or S-wave velocity *β*. To discretize the wave equation, we limit ourselves to a discrete number of points in time and space. Thus, instead of *x* as a continuous variable, we consider *grid points*

$$x_0 = 0,\ x_1 = \Delta x,\ x_2 = 2\Delta x,\ \ldots,\ x_k = k\Delta x,\ \ldots, \tag{6.95}$$

with constant *grid spacing* Δ*x*. Similarly, we take grid points in time

$$t_0 = 0,\ t_1 = \Delta t,\ t_2 = 2\Delta t,\ \ldots,\ t_i = i\Delta t,\ \ldots, \tag{6.96}$$

with constant *time step* Δt. We may now approximate the partial derivatives in Eq. (6.94) by difference quotients evaluated at the grid points. For instance, using grid points x_k and $x_{k+1} = x_k + \Delta x$, we can approximate a first derivative with respect to x (Fig. 6.37):

$$\frac{\partial u(x_k, t_i)}{\partial x} \approx \frac{1}{\Delta x}\left[u(x_{k+1}, t_i) - u(x_k, t_i)\right]. \tag{6.97}$$

To obtain an approximation of the second derivative, we simply repeat this procedure, now using grid points x_{k-1} and x_k as in Fig. 6.36:

$$\frac{\partial^2 u(x_k, t_i)}{\partial x^2} \approx \frac{1}{\Delta x}\left[\frac{\partial u(x_k, t_i)}{\partial x} - \frac{\partial u(x_{k-1}, t_i)}{\partial x}\right]. \tag{6.98}$$

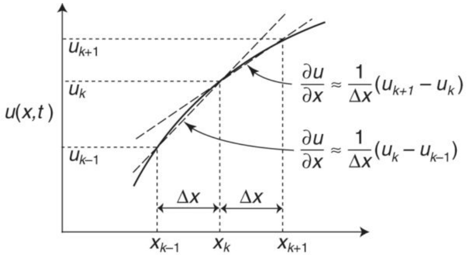

u_{k+1}
u_k
$u(x,t)$
u_{k-1}
$\frac{\partial u}{\partial x} \approx \frac{1}{\Delta x}(u_{k+1} - u_k)$
$\frac{\partial u}{\partial x} \approx \frac{1}{\Delta x}(u_k - u_{k-1})$
Δx
Δx
x_{k-1}
x_k
x_{k+1}

Fig. 6.37 Illustration of two possible finite-difference approximations of a derivative. One uses grid points x_{k+1} and x_k, the other uses x_k and x_{k-1}. Both are employed to obtain the finite-difference approximation of the second derivative in Eq. (6.99).

Substituting Eq. (6.97) for $\partial u(x_k,t_i)/\partial x$ and for the neighboring grid point, $\partial u(x_{k-1},t_i)/\partial x$, into Eq. (6.98), we obtain a finite-difference approximation of the second derivative,

$$\frac{\partial^2 u(x_k, t_i)}{\partial x^2} \approx \frac{1}{\Delta x^2}\left[u(x_{k+1}, t_i) - 2u(x_k, t_i) + u(x_{k-1}, t_i)\right]. \tag{6.99}$$

The derivation of Eq. (6.99) can be repeated to approximate the second *time* derivative of the wavefield. Instead of neighboring grid points in space, we take neighboring grid points in time,

$$\frac{\partial^2 u(x_k, t_i)}{\partial t^2} \approx \frac{1}{\Delta t^2}\left[u(x_k, t_{i+1}) - 2u(x_k, t_i) + u(x_k, t_{i-1})\right]. \tag{6.100}$$

We may now substitute the approximations, Eqs. (6.99) and (6.100), into Eq. (6.94). Using the abbreviated notation $u_{k,i}$ for the wavefield at position x_k and time t_i, we find

$$\frac{1}{\Delta t^2}(u_{k,\, i+1} - 2u_{k,\, i} + u_{k,\, i-1}) \approx \frac{c_k^2}{\Delta x^2}(u_{k+1,\, i} - 2u_{k,\, i} + u_{k-1,\, i}). \tag{6.101}$$

Equation (6.101) is a discrete version of the wave equation. All derivatives have been replaced by sums and differences that can be calculated using a computer. To see the practical usefulness of Eq. (6.101), we rearrange terms such that $u_{k,i+1}$ appears on the left-hand side:

$$u_{k,\, i+1} = 2u_{k,\, i} - u_{k,\, i-1} + \frac{c_k^2 \Delta t^2}{\Delta x^2}(u_{k+1,\, i} - 2u_{k,\, i} + u_{k-1,\, i}). \tag{6.102}$$

Equation (6.102) is in fact a convenient numerical recipe: Letting t_i be the current time, the wavefield $u_{k,i+1}$ for the future time step t_{i+1} can be computed from the current wavefield $u_{k,i}$ and the past wavefield $u_{k,i-1}$. Once the future wavefield $u_{k,i+1}$ is computed, we repeat the same procedure: Using $u_{k,i+1}$ and $u_{k,i}$, we compute $u_{k,i+2}$, and so on. Thus, we obtain a numerical solution *iteratively* by time-stepping.

An example of the practical implementation of Eq. (6.102) is shown in Fig. 6.38. The velocity distribution (Fig. 6.38a) is heterogeneous, including a gradual velocity increase around $x = -$

200 km, as well as three velocity discontinuities, labeled d_1, d_2, and d_3. At t = 0 s, the wave starts to propagate away from its initial position at x= 0 km (Fig. 6.38b). The initial wave with amplitude 2 mm splits into two waves with half the amplitude, one propagating to the left and one propagating to the right. They are labeled w_L and w_R, respectively, in Fig. 6.38c. When w_L and w_R reach a discontinuity, a reflected wave is produced. For example, $w_{L,R1}$ is the reflection caused by the interaction of w_L with discontinuity d_1. It travels to the right. Similarly, the leftward traveling reflection $w_{R,L3}$ results from the interaction of w_R with discontinuity d_3. At each discontinuity, the amplitude of the transmitted waves w_L and w_R is reduced.

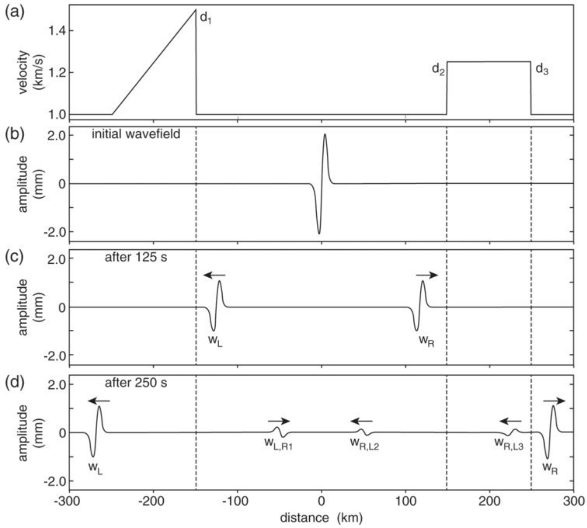


Fig. 6.38 Wavefield computed numerically for a heterogeneous velocity distribution using the finite-difference approximation of Eq. (6.102). The grid spacing for this example is Δx =0.5 km, and the time step is Δt =0.25 s. (a) Velocity $c(x)$ along a 600 km long line, including three discontinuities, d_1, d_2, and d_3. (b) Wavefield snapshot at the initial time t =0 s. (c) At t =125 s, one wave, labeled w_L, has propagated leftward to x = –125 km. Another wave, w_R, propagating to the right, has reached x = 125 km. (d) After 250 s, w_L and w_R have interacted with several discontinuities, marked by dashed lines. The result are reflected waves with lower amplitude, labeled $w_{L,R1}$, $w_{R,L2}$, and $w_{R,L3}$.

The advantage of this numerical solution is that the complete wavefield, including all multiply reflected waves, is computed automatically for velocity distributions $c(x)$ that may be very heterogeneous. This eliminates the need to track countless individual reflected and transmitted waves, as we did in Section 6.3.5.

However, as Eqs. (6.97) to (6.100) already suggest, numerical solutions are generally not quite perfect. They involve approximations, the quality of which depends on the choice of the grid spacing Δx and the time step Δt, and also on the distribution of velocity $c(x)$. The *finite-difference* approximation in Eq. (6.97) improves when the grid spacing Δx decreases, and it would approach the correct derivative $\partial u/\partial x$ for Δx tending toward 0. However, decreasing Δx also increases the number of grid points needed to cover a certain distance, and therefore the amount of computations. For an infinitesimally small Δx we have infinitely many grid points, and infinitely many mathematical operations are needed to advance the wavefield from the current time t_i to the future time t_{i+1}. Clearly, this is not possible, and so we must choose Δx such that our computer can handle the calculations. Similar arguments hold for the time step Δt. A very small Δt improves the finite-difference approximation at the expense of having to compute many time steps in order to reach a certain time. Conversely, a Δt that is too large may dramatically deteriorate the quality of the numerical solution.

6.3.7.2 Alternative Methods and Modern Developments

Using the approximations of Eqs. (6.97) and (6.98) for derivatives in the y- and z-directions, the finite-difference method can be extended to the simulation of wave propagation in two and three dimensions. The resulting scheme is easily implemented in a computer program, but it is far from being the only method to discretize the wave equation. For instance, a larger number of neighboring grid points around x_k may be used to obtain a more accurate approximation of a derivative for a given Δx. Also, the approximation used for the time derivatives need not be the same as for the space derivatives.

A different class of numerical methods – *finite-element* analysis – discretizes the wave equation by subdividing the Earth into small subvolumes, or *elements*, within which the material properties are nearly constant. The wave equation is then solved with high accuracy for the individual elements, and finally the solution is patched together by forcing the wavefield to be continuous across the element boundaries. Finite-element methods allow us to account very accurately for many complexities of the Earth. An example is shown in Fig. 6.39. The seismic source, indicated by a star, is located at the surface of the ocean, which is fluid. From there, a pressure wave travels to the ocean bottom, which is the interface to the underlying solid Earth, where both P- and S-waves exist. Instead of being perfectly circular, the P- and S-wavefronts, shown as thin curved lines in Fig. 6.39, are distorted or even broken into smaller waves. This complexity is the result of repeated reflections and diffractions in the complicated underground structure, which includes a fault and multiple folded sediment layers. At the ocean bottom, the initial pressure wave is also reflected back upwards to the ocean surface. From there it reflects downwards again, and then also enters into the solid Earth as another set of P- and S-waves, labeled P_{surf} and S_{surf} in Fig. 6.39. This process of surface and ocean-bottom reflection repeats, each time generating a new pair of P- and S-waves with lower amplitudes. The result is a complicated, yet very accurately computed, seismic wavefield.

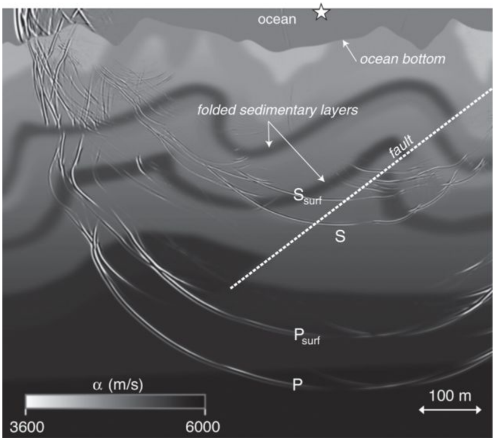

ocean
ocean bottom
folded sedimentary layers
fault
S_{surf}
S
P_{surf}
P
α (m/s)
3600
6000
100 m

Fig. 6.39 Snapshot of a seismic wavefield superimposed on the P-wave velocity distribution in the background, which features a dipping fault and several folded sediment layers. The position of the source within the overlying ocean layer is indicated by a star. In addition to the direct P- and S-wave, there are waves that were reflected at the ocean bottom and then at the ocean surface, before being transmitted downwards into the solid Earth. These are labeled P_{surf} and S_{surf} (courtesy of Michael Afanasiev).

6.4 Ambient Seismic Waves

6.4.1 Introduction

In the late 1860s, the Italian priest Timoteo Bertelli studied the movement of a pendulum installed on the wall of the college where he taught natural sciences. He observed a phenomenon that he could trace back in historic records at least to the year 1643: The pendulum moved spontaneously, without any obvious external excitation! After thousands of tedious and careful experiments he found that the microscopic movements, which he called *microseisms*, correlated with barometric depressions, especially in winter. Bertelli's pendulum functioned as a bad weather detector!

More than 150 years later, we know that the Earth vibrates continuously under the influence of a wide range of forces that act on its surface. They include long-period variations of atmospheric pressure, daily temperature changes that cause thermal expansion and contraction in the shallow subsurface, the interaction of waves with the coast line and the ocean floor, roots that transmit the movement of trees in the wind into the Earth, but also human and industrial activity (Fig. 6.40).

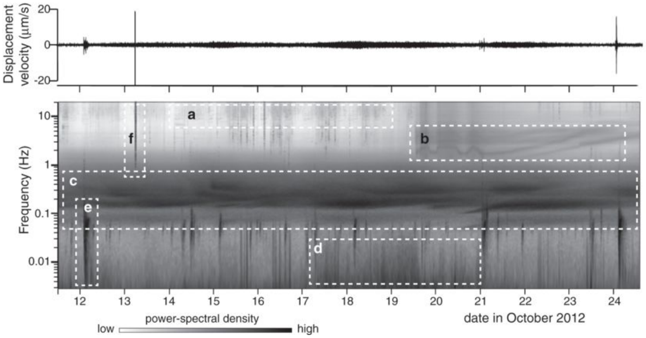


Fig. 6.40 Vertical-component displacement velocity at a seismic station in the northern Pyrenees during two weeks in October 2012. The power-spectral density plotted below the time-series measures the signal amplitude as a function of time and frequency. (a) High-frequency day–night variations of anthropogenic noise. (b) Water discharge of the Aragon river following heavy rainfall. (c) Microseismic noise related to interactions of ocean waves with the

solid Earth. (d) Low-frequency noise referred to as *hum* and related to the remnants of hurricane Rafael in the North Atlantic. (e) Magnitude 6.7 earthquake near Papua New Guinea at a distance of around 14,000 km. (f) Magnitude 3.8 earthquake near the French town of Lourdes at around 60 km distance (courtesy of Jordi Diaz).

These *ambient sources* are distinctly different from transient sources such as earthquakes or explosions. They may act over a long period, cover a large area, and vary almost randomly in time and space. Consequently, the *ambient wavefield* does not consist of well-defined P-wave, S-wave, and surface wave packets that are easily identifiable in a seismic recording. Instead, the ambient wavefield is a continuous superposition of all possible wave types.

The ambient wavefield is superimposed onto the wavefields related to transient sources. It therefore tends to obscure information that seismologists have traditionally been interested in: travel times and amplitudes of clearly visible seismic phases. From this perspective, it is not surprising that the ambient wavefield has often been classified as *ambient noise*. Today we know that the ambient wavefield is a valuable carrier of information on Earth structure. This information can be extracted using methods collectively known as *seismic interferometry* (Section 6.4.3). Nevertheless, the term ambient noise stubbornly persists, maybe to emphasize the apparent paradox that signals historically classified as nuisance may actually be very useful. In the following, we will use the terms ambient field and ambient noise interchangeably.

6.4.2 Nature and Sources of Ambient Noise

6.4.2.1 Phenomenology of Ambient Noise Recordings

Individual recordings of ambient seismic noise generally look random (Fig. 6.41). The lack of structure excludes the application of seismic analysis tools that rely on the identification of specific phases and their properties, e.g., travel times and amplitudes.

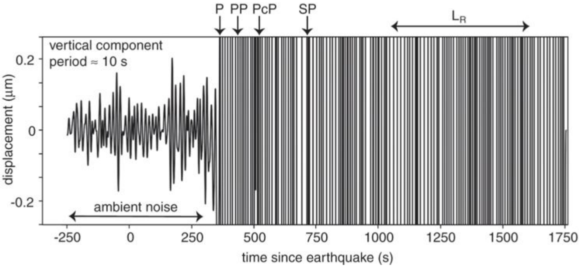


Fig. 6.41 Illustration of ambient noise prior to the arrival of a P-wave. The seismogram is exactly the same as in Fig. 6.1a but the plot is stretched vertically by a factor of 1000 to make the ambient noise visible.

The amplitude of ambient noise is strongly frequency-dependent. Especially in densely populated areas, anthropogenic noise is significant at frequencies above 1 Hz (Fig. 6.40a). The strongest ambient noise is almost always recorded at frequencies of around 0.08 and 0.16 Hz. These frequencies of maximum noise amplitude are referred to as the *primary* and *secondary microseismic peaks*, respectively (Fig. 6.40c). Between frequencies of around 0.03 and 0.01 Hz, ambient noise is typically low. This is a convenient coincidence, as most surface wave analyses are performed in this frequency band. At even lower frequencies, below 10 mHz, the Earth performs slow free oscillations that affect the planet as a whole. This kind of ambient noise is known as the *Earth's hum* (Fig. 6.40d).

Though ambient noise on individual records appears rather structureless, it has remarkable spatial coherence; ambient noise recorded at one receiver is similar to what is recorded by another one located nearby. This coherence distinguishes ambient from instrumental noise. The latter would indeed be different at all receivers. The ambient field is coherent because the laws of physics prevent complete randomness. As any other seismic disturbance, the ambient field is a wave that propagates at a velocity c determined by the material properties of the Earth, and for a given frequency f it has a characteristic wavelength $\lambda = c/f$.

With the advent of dense receiver networks, the spatial coherence of the ambient field has become visible. This is illustrated in Fig. 6.42, which shows vertical displacement velocity recorded by a large number of seismometers located at 70 m water depth in the North Sea, just above the Valhall oil field. The average receiver spacing of about 50 m provides a detailed snapshot of the wavefield, which contains wave packets with a wavelength roughly estimated at $\lambda = 1$ km. Multiplying this estimate with the center frequency of the recording, $f = 0.3$ Hz, yields a velocity of $c = \lambda f = 0.3$ km s^{-1}, which closely corresponds to the surface wave velocity measured with active source (explosion) experiments. This example illustrates that already a visual analysis of the ambient field contains information about the Earth's interior. While receiver spacing is not generally as dense as directly above a hydrocarbon reservoir, the tools of noise interferometry still allow us to exploit ambient field coherence at receivers that are hundreds or thousands of kilometers apart (Section 6.4.3).

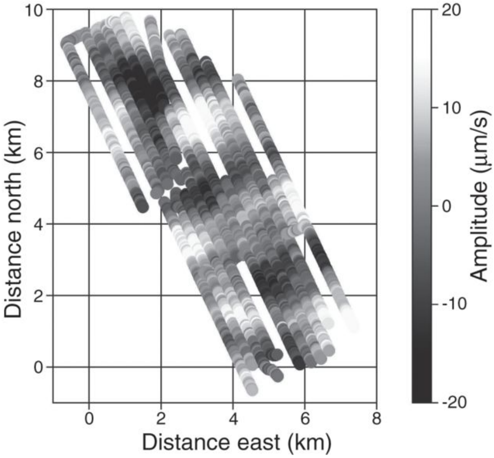


Fig. 6.42 Vertical-component displacement velocity of the ambient wavefield recorded at UTC 0:17:00 on December 24, 2010 above the Valhall oil field at 70 m water depth in the North Sea. Each filled circle represents a receiver. The frequency band is 0.25–0.35 Hz, close to the secondary microseismic peak (courtesy of Sjoerd de Ridder, based on data by Aker BP AS and Pandion Energy AS).

6.4.2.2 Physical Origin of Ambient Noise

The physical processes responsible for the generation of ambient noise are diverse, and the boundaries between the frequency bands where they are active are not always clear-cut. While ambient noise at frequencies roughly above 1 Hz is commonly attributed to human activity, the primary and secondary microseismic peaks can be explained by a specific type of ocean wave known as *swell*. Unlike small local wind waves with few-meter wavelengths, swell is the result of large and long-lasting weather systems. On average, swell has a period of around 12 s (frequency 0.08 Hz), a wavelength on the order of 100 m, and wave heights of several meters. This makes it popular with surfers. In the shallow continental shelf regions, swell exerts pressure on the ocean floor, thereby generating seismic waves, mostly Rayleigh and Love waves, at the same period, i.e., approximately 12 s. Similarly, swell interacts with the coastline, where additional ocean wave energy at around 12 s period is transformed into seismic waves.

Secondary microseisms are produced when swell from one weather system interferes with swell propagating in the opposite direction, coming from another weather system. The resulting pressure fluctuations reach greater depth than each of the individual waves, and they have half the period of the original swell, i.e., around 6 s (frequency 0.16 Hz). Therefore, secondary microseisms are predominantly generated in deeper water.

The physical mechanism behind the long-period hum is more debated, but the influence of large weather systems, and tropical cyclones in particular, is undisputed. Most likely, ocean waves caused by such extreme weather interact directly with the ocean floor, similar to the mechanism responsible for the primary microseisms.

6.4.3 Noise Interferometry

To understand how useful information can be extracted systematically from recordings of seemingly random ambient noise, we consider two receiver positions **x** and **y** at a distance L from each other (Fig. 6.43). For simplicity, we assume that the medium is homogeneous with velocity c, so that a wave takes the time L/c to travel from **x** to **y**.

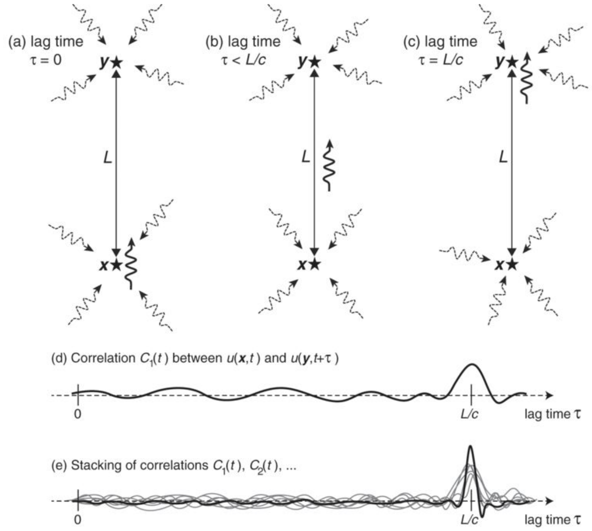

(a) lag time
τ = 0
y
L
x
(b) lag time
τ < L/c
y
L
x
(c) lag time
τ = L/c
y
L
x
(d) Correlation C₁(t) between u(x,t) and u(y,t+τ)
0
L/c
lag time τ
(e) Stacking of correlations C₁(t), C₂(t), ...
0
L/c
lag time τ

Fig. 6.43 Schematic illustration of ambient noise correlation. (a) Receivers at **x** and **y**, separated by distance L. Wavy arrows indicate wave packets reaching the receivers from random directions. The black wave packet travels toward receiver **y**. (b) After some time, another set of wave packets reaches the receivers, and the black one approaches **y**. (c) After L/c seconds, the black wave packet has reached **y**. As a consequence, the time shifted $u(\mathbf{y},t+ L/c)$ is similar to $u(\mathbf{x},t)$. (d) This similarity is expressed in the correlation function $C_1(\tau)$, which has a slight peak for lag times around $\tau = L/c$. (e) Stacking many correlation functions (gray thin curves) with different start times enhances the peak and makes the stacked correlation function (thick black curve) appear like a wave emitted from a virtual source at **x** and received at **y** after the travel time L/c.

At some time t_1, the ambient noise recording at **x**, denoted by

$u(\mathbf{x},t_1)$, is a superposition of wave packets arriving from random directions (light-gray wavy arrows in Fig. 6.43a). One of these many wave packets travels in a direction approximately oriented toward **y** (bold wavy arrow in Fig. 6.43a). At the same time t_1, the recording $u(\mathbf{y},t_1)$ at position **y** is composed of another set of wave packets arriving randomly from various directions.

Box 6.1 Correlation and the Similarity of Time-Series

A frequent task in the analysis of time-series is the quantification of their similarity, which can be achieved with the help of *cross-correlation* or simply *correlation*. To illustrate the concept, consider two time-series, e.g., two seismograms $u(\mathbf{x},t)$ and $u(\mathbf{y},t)$ recorded at positions defined by vectors **x** and **y**. The time-series start at time $t = t_1$ and end at time $t = t_1 + \Delta$ (Fig. B6.1.1a). Their correlation is defined as

$$C_1 = \frac{1}{\Delta} \int_{t_1}^{t_1+\Delta} u(\mathbf{x},t)u(\mathbf{y},t)\mathrm{d}t. \tag{1}$$

To see why correlation quantifies similarity, we approximate the integral in Eq. (1) by a sum,

$$C_1 \approx \sum_k u(\mathbf{x}, t_1 + k\ \mathrm{d}t)\ u(\mathbf{y}, t_1 + k\ \mathrm{d}t)\mathrm{d}t, \tag{2}$$

where dt is a small time increment. When the time-series are dissimilar, as in Fig. B6.1.1a, the signs of the individual products $u(\mathbf{x},t_1 + k\mathrm{d}t)\ u(\mathbf{y},t_1 + k\mathrm{d}t)$ are sometimes positive and sometimes negative. As a consequence, the contributions tend to cancel each other, and the numerical value of the correlation

C_1 is small.

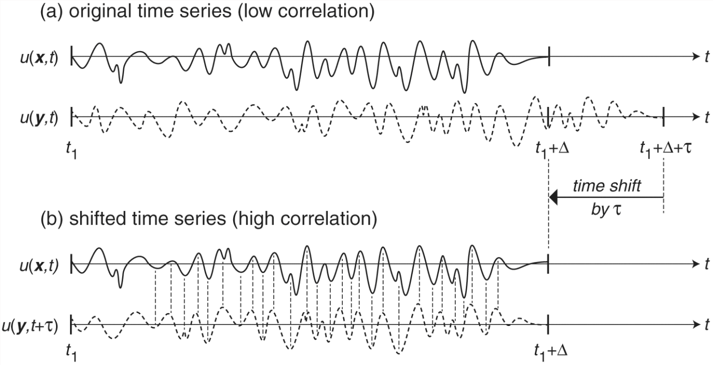

(a) original time series (low correlation)
u(x,t)
t
u(y,t)
t
t_1
$t_1+\Delta$
$t_1+\Delta+\tau$
time shift
by τ
(b) shifted time series (high correlation)
u(x,t)
t
u(y,t+τ)
t
t_1
$t_1+\Delta$

Fig. B6.1.1 Schematic illustration of time-series similarity. (a) The peaks and troughs of the time series $u(\boldsymbol{x},t)$ and $u(\mathbf{y},t)$ mostly do not coincide. This lack of similarity translates into a low value of the correlation C_1 (Eq. 1). (b) Shifting the time-series $u(\mathbf{y},t)$ to $u(\mathbf{y},t+\tau)$ aligns its peaks and troughs almost perfectly with those of $u(\boldsymbol{x},t)$. As a consequence, the correlation of $u(\mathbf{y},t+\tau)$ and $u(\boldsymbol{x},t)$, denoted by $C_1(\tau)$, is larger.

Shifting the time-series $u(\mathbf{y},t)$ by some time τ to $u(\mathbf{y},t+\tau)$ may improve the alignment, i.e., the similarity, with time-series $u(\mathbf{x},t)$. In fact, as shown in Fig. B6.1.1b, the peaks and troughs of $u(\mathbf{y},t+\tau)$ and $u(\mathbf{x},t)$ occur at similar times. Computing the time-shifted correlation

$$\begin{aligned} C_1(\tau) &= \frac{1}{\Delta}\int_{t_1}^{t_1+\Delta} u(\mathbf{x},t)\ u(\mathbf{y},t+\tau)\ \mathrm{d}t \\ &\approx \sum_k u(\mathbf{x},t_1+k\ \mathrm{d}t)\ u(\mathbf{y},t_1+\tau+k\ \mathrm{d}t)\ \mathrm{d}t \end{aligned} \tag{3}$$

gives individual contributions $u(\mathbf{x},t_1+k\mathrm{d}t)\ u(\mathbf{y},t_1+\tau+k\mathrm{d}t)$ that all have a positive sign. Thus, instead of canceling each other, the contributions add up, leading to a large value of the correlation, which indicates similarity.

To find the time shifts or *lag times* at which two time-series are most similar or dissimilar, τ is treated as a variable. The similarity is largest where the correlation $C_1(\tau)$ attains a maximum, i.e., where its first derivative with respect to τ is zero and the second derivative is negative.

Since the wave packets arriving at **x** are different from those arriving at **y**, the recordings at the two positions are rather different within the time interval from t_1 to $t_1 + \Delta$. This means that the correlation

$$C_1 = \frac{1}{\Delta} \int_{t_1}^{t_1+\Delta} u(\mathbf{x}, t) u(\mathbf{y}, t) \mathrm{d}t, \tag{6.103}$$

which measures similarity between $u(\mathbf{x},t)$ and $u(\mathbf{y},t)$, takes a small value (Box 6.1). As time progresses, the wave packets continue to propagate, and new ones arrive from random directions at **x** and **y**. The wave packet that was oriented along the receiver line continues its journey toward **y** (bold wavy arrow in Fig. 6.43b). After time L/c, this wave packet finally arrives at **y** (Fig. 6.43c). As a consequence, the time-shifted recording $u(\mathbf{y},t+ L/c)$ is more similar to the recording $u(\mathbf{x},t)$ made at position **x** and L/c seconds earlier. This translates into a correlation

$$C_1(\tau) = \frac{1}{\Delta} \int_{t_1}^{t_1+\Delta} u(\mathbf{x}, t) u(\mathbf{y}, t + \tau) \mathrm{d}t, \tag{6.104}$$

which is larger than $C_1(\tau = 0)$ in Eq. (6.103). Considering the correlation function as a function of the time shift or lag time τ, we therefore expect to observe a slight peak around $\tau = L/c$ (Fig. 6.43d). Since other random wave packets also arrive at $\boldsymbol{y}$ around time $t_1 + L/$

c, the correlation at $\tau = L/c$ is usually not perfect. However, it can be improved substantially by summing or *stacking* correlations for many successive times $t_2 = t_1 + \Delta$, $t_3 = t_2 + \Delta$, ...

$$C(\tau) = \sum_n C_n(\tau). \tag{6.105}$$

The stacking suppresses the contributions of wave packets that arrive from random directions, and it enhances the contribution of wave packets traveling from **x** to **y** within the time L/c (Fig. 6.43e). Since the stacked correlation function $C(\tau)$ differs significantly from zero only for $\tau = L/c$, it has the appearance of an individual wave packet emitted by a *virtual source* at receiver **x** and recorded at **y**.

This result is remarkable for many reasons. Most importantly, noise has been transformed into a signal that resembles a wave excited by an earthquake or an explosion. By measuring the travel time of that signal in the correlation function, the propagation velocity c between the receivers is inferred, which could not have been done by analyzing the noise recordings individually. Furthermore, there is now a virtual source at a position where no real source is actually located.

While we have demonstrated ambient noise interferometry with a simplified intuitive model, the concept can be shown mathematically to work for the Earth in three dimensions. Ambient noise interferometry allows us to place virtual sources in regions where real sources such as earthquakes or explosions are not available. The method can be applied on all scales, ranging from reservoirs a few kilometers in size (Fig. 6.42) to the whole Earth (Fig. 6.44).

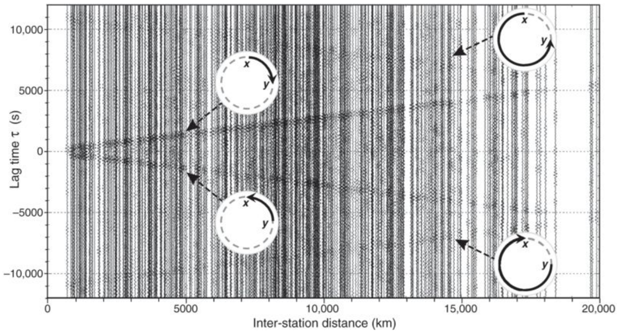


Fig. 6.44 Ambient noise correlations for stations distributed around the globe. The maximum inter-station distance of ∼20,000 km corresponds to half the circumference of the Earth. The visible wave packets mostly correspond to Rayleigh waves traveling around the globe. Circles and arrows indicate the direction of wave propagation. Positive lag times τ correspond to waves traveling from $\boldsymbol{x}$ to $\boldsymbol{y}$, either along the minor arc (shorter distance) or the major arc (longer distance). Similarly, negative times correspond to waves traveling in the opposite direction from $\boldsymbol{y}$ to $\boldsymbol{x}$ (courtesy of Laura Ermert).

6.5 Seismometry

6.5.1 Introduction

Seismology is a data-driven and data-rich science that is based on the recording of seismic waves by receivers, called *seismometers*. In fact, the beginnings of seismology as a quantitative science date back to the second half of the nineteenth century, when the first seismometers were developed by Robert Mallet, John Milne, and others. The overwhelming majority of seismometers developed since the days of Mallet and Milne exploit the principle of inertia. If a heavy mass is only loosely coupled to the ground (for example, by suspending it from a wire like a pendulum as in Fig. 6.45), the motion of the Earth caused by a seismic wave is only partly transferred to the mass. While the ground vibrates, the inertia of the heavy mass assures that it does not move as much. The seismometer records the relative motion between the mass and the ground.

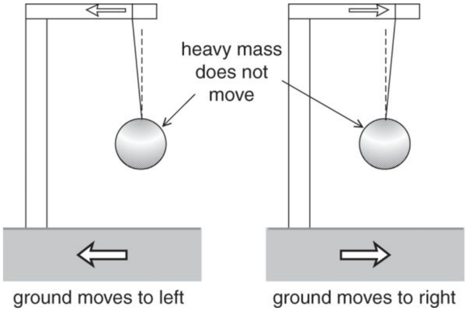


Fig. 6.45 The principle of the inertial seismometer. Because of its inertia, a suspended heavy mass remains almost stationary when the ground and suspension move to the left or to the right.

Early seismometers were undamped and reacted only to a limited band of seismic frequencies. Seismic waves with inappropriate frequencies were barely recorded at all, but strong waves could set the instrument into resonant vibration. In 1903, the German seismologist Emil Wiechert substantially increased the measurement accuracy by damping the instrument. These early instruments relied on mechanical levers for amplification and recording signals on smoked paper. This made them both bulky and heavy, which severely restricted their application.

A major technological improvement was achieved in 1906, when Prince Boris Galitzin of Russia introduced the electromagnetic seismometer, which allowed galvanometric recording on photographic paper. This electrical method had the great advantage that the recorder could now be separated from the seismometer.

Modern seismometers are highly complex instruments, capable of measuring ground displacement in the nanometer range. They contain electronic feedback systems which allow us to record a broad spectrum of frequencies, including the lowest-frequency free oscillations (below 1 mHz) and P-waves traveling all the way through the Earth (above 1 Hz). Here, we will mostly limit ourselves to a description of the basic principles of an inertial seismometer.

6.5.2 Principle of the Inertial Seismometer

Most inertial seismometers employ variations on the pendulum principle (Fig. 6.45) or mass-spring systems that vibrate when subject to ground motion. Seismometers may be categorized into vertical- and horizontal-motion instruments, depending on how the pendulum or the spring are oriented. Some modern instruments are constructed so as to record simultaneously three orthogonal components of motion.

6.5.2.1 Vertical-Motion Seismometer

In the mechanical type of vertical-motion seismometer (Fig. 6.46a), a large mass is mounted on a horizontal bar hinged at a pivot so that it can move only in the vertical plane. The bar is held in a horizontal position by a weak spring. This assures a loose coupling between the mass and the housing, which is connected rigidly to the ground. Vertical ground motion, as sensed during the passage of a seismic wave, is transmitted to the housing but not to the inertial mass, which remains stationary.

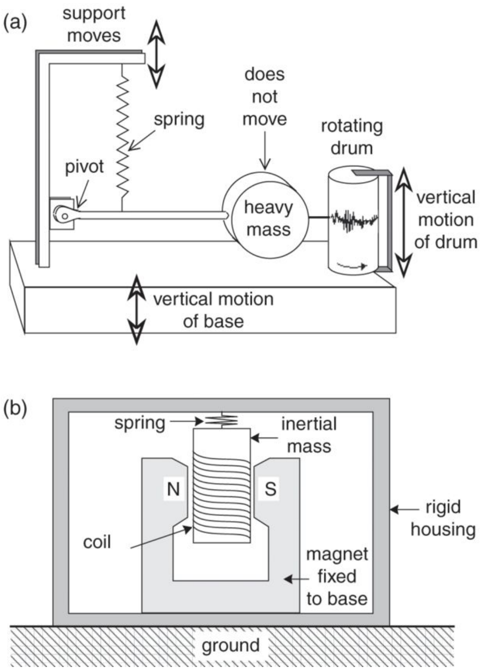

(a)
support moves
spring
pivot
does not move
heavy mass
rotating drum
vertical motion of drum
vertical motion of base
(b)
spring
inertial mass
N
S
coil
rigid housing
magnet fixed to base
ground

Fig. 6.46 Schematic diagrams illustrating the principle of operation of the vertical-motion seismometer: (a) mechanical pendulum type; (b) electromagnetic, moving-coil type.

Source: A.N. Strahler, *The Earth Sciences*, 681 pp., New York: Harper and Row, 1963.

An alternative design responds to the relative motion between a magnet and a coil of wire. One of these members is fixed to the housing of the instrument and thereby to the Earth. The other is suspended by a spring and forms the inertial member. Two basic designs are possible. In the moving-magnet type, the coil is fixed to the housing and the magnet is inertial. In the moving-coil type the roles are reversed (Fig. 6.46b). A coil of wire fixed to the inertial mass is suspended between the poles of a strong magnet, which in turn is fixed to the ground by the rigid housing. Any motion of the coil within the magnetic field induces a voltage in the coil proportional to the rate of change of magnetic flux. During a seismic arrival the vibration of the ground relative to the mass is converted to an electrical voltage by induction in the coil. The voltage is amplified and transmitted through an electrical circuit to the recorder. In a similar fashion, the movement of the mass in the pendulum design (Fig. 6.46a) can be converted into a voltage and stored digitally.

6.5.2.2 Horizontal-Motion Seismometer

The principle of the mechanical type of horizontal-motion seismometer is similar to that of the vertical-motion instrument. As before, the inertial mass is mounted on a horizontal bar, but the fulcrum is now hinged almost vertically so that the mass is confined to swing sideways in a nearly horizontal plane (Fig. 6.47). The behavior of the system is similar to that of a gate when its hinges are slightly out of vertical alignment. If the hinge axis is tilted slightly forward, the stable position of the gate is where its center of mass is at its lowest point. In any displacement of the gate, the restoring gravitational forces try to return it to this stable position. Similarly, the horizontal-motion seismometer swings about its equilibrium position like a horizontal pendulum (in fact it is the housing of the instrument that moves and not the inertial mass). As in the vertical-motion seismometer, the motion of the mass is converted to voltage via electromagnetic induction, and then stored digitally. The voltage recordings can be calibrated to represent displacement velocity of the ground.

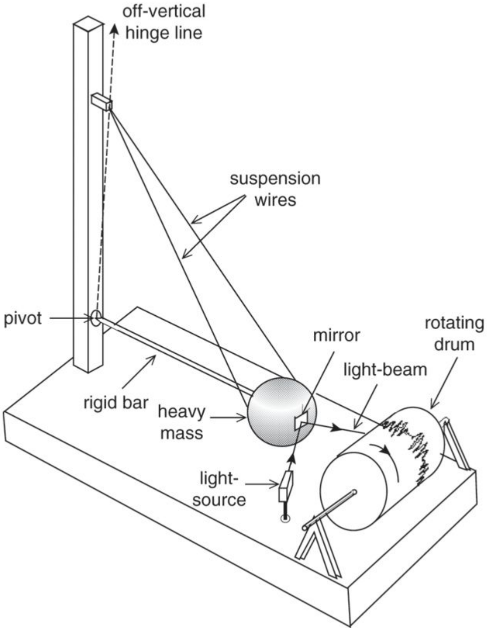

off-vertical
hinge line
suspension
wires
pivot
mirror
rotating
drum
light-beam
rigid bar
heavy
mass
light-
source

Fig. 6.47 Schematic design of the pendulum type of horizontal-motion seismometer. As for the vertical-motion type, the motion detection in modern instruments is made electromagnetically.

Source: A.N. Strahler, *The Earth Sciences*, 681 pp., New York: Harper and Row, 1963.

6.5.3 The Equation of the Seismometer

Inertial seismometers for recording horizontal and vertical ground motion function on the pendulum principle. When the instrument frame is displaced from its equilibrium position relative to the inertial mass, a restoring force arises that is, to first order, proportional to the displacement. Let the vertical or horizontal displacement, dependent on the type of seismometer, be u and the restoring force $-ku$, and let the corresponding displacement of the ground be q. The total displacement of the inertial mass M is then $(u + q)$, and the equation of motion is

$$M\frac{\partial^2}{\partial t^2}(u+q) = -ku.$$

(6.106)

We now divide throughout by M, write $k/M = \omega_0^2$, and after rearranging the equation, we get the familiar equation of forced simple harmonic motion,

$$\frac{\partial^2 u}{\partial t^2} + \omega_0^2 u = -\frac{\partial^2 q}{\partial t^2}.$$

(6.107)

In this equation, ω_0 is the *natural frequency* (or *resonant frequency*) of the instrument. For a ground motion with this frequency, the seismometer would execute large uncontrolled vibrations and the seismic signal could not be accurately recorded. To get around this problem, the seismometer motion is damped by providing a velocity-dependent force that opposes the motion. A damping term enters into the equation of motion, which becomes

$$\frac{\partial^2 u}{\partial t^2} + 2\lambda\omega_0\frac{\partial u}{\partial t} + \omega_0^2 u = -\frac{\partial^2 q}{\partial t^2}.$$

(6.108)

The constant λ in this equation is called the *damping factor* of the instrument. It plays an important role in determining how the seismometer responds to a seismic wave.

A seismic signal is generally composed of numerous superposed harmonic vibrations with different frequencies. We can determine how a seismometer with natural frequency ω_0 responds to a seismic signal with any frequency ω by solving Eq. (6.108) with $q = A\cos\omega t$ (Box 6.2). Here, A is the magnified amplitude of the ground motion, equal to the true ground motion multiplied by a magnification factor that depends on the sensitivity of the instrument. Let the displacement u recorded by the seismometer be

$$u = U\cos(\omega t - \Delta),$$

(6.109)

where U is the amplitude of the recorded signal and Δ is the phase difference between the record and the ground motion. As derived in Box 6.2, the phase lag Δ is given by

$$\Delta = \tan^{-1}\left(\frac{2\lambda\omega\omega_0}{\omega_0^2 - \omega^2}\right).$$

(6.110)

The solution of the equation of motion (Eq. 6.108) gives the displacement u on the seismic record as

$$u = \frac{A\omega^2}{\sqrt{(\omega_0^2 - \omega^2) + 4\lambda^2\omega^2\omega_0^2}}\cos(\omega t - \Delta).$$

(6.111)

Box 6.2 The Seismometer Equation

Equation (6.108) is the damped equation of motion for the signal u recorded by a seismometer with natural frequency ω_0 and damping factor λ, when the ground displacement is q. Let the ground displacement vary with time according to $q = A\cos\omega t$ and let the recorded seismic signal be $u = U\cos(\omega t - \Delta)$. Substituting in Eq. (6.108), we get the following:

$$-\omega^2 U \cos(\omega t - \Delta) - 2\lambda\omega\omega_0 U \sin(\omega t - \Delta) + \omega_0^2 U \cos(\omega t - \Delta) = A\omega^2 \cos\omega t,$$

(1)

i.e.,

$$U[(\omega_0^2 - \omega^2)\cos(\omega t - \Delta) - 2\lambda\omega\omega_0 \sin(\omega t - \Delta)] = A\omega^2 \cos\omega t,$$

(2)

If we now write

$$(\omega_0^2 - \omega^2) = R\cos\varphi \quad \text{and} \quad 2\lambda\omega\omega_0 = R\sin\varphi,$$

(3)

where

$$R = [(\omega_0^2 - \omega^2)^2 + 4\lambda^2\omega^2\omega_0^2]^{1/2} \quad \text{and} \quad \tan\varphi = \left(\frac{2\lambda\omega\omega_0}{\omega_0^2 - \omega^2}\right).$$

(4)

These equations reduce to

$$U[R\cos\varphi\cos(\omega t - \Delta) - R\sin\varphi\sin(\omega t - \Delta)] = A\omega^2 \cos\omega t,$$

(5)

$$UR\cos(\omega t-\Delta+\varphi)=A\omega^2\cos\omega t.$$

(6)

Complex numbers (see Box 4.6) allow a simple solution to this equation. The function $\cos\theta$ is the real part of the complex number $\mathrm{e}^{\mathrm{i}\theta}$ (i.e., $\cos\theta=\mathrm{Re}\{\mathrm{e}^{\mathrm{i}\theta}\}$). Therefore,

$$UR\mathrm{Re}\{\mathrm{e}^{(\mathrm{i}(\omega t-\Delta+\varphi))}\}=A\omega^2\mathrm{Re}\{\mathrm{e}^{(\mathrm{i}\omega t)}\},$$

(7)

$$\begin{aligned}U&=\frac{A\omega^2}{R}\mathrm{Re}\{\mathrm{e}^{(\mathrm{i}\omega t)}\mathrm{e}^{-(\mathrm{i}(\omega t-\Delta+\varphi))}\}=\frac{A\omega^2}{R}\mathrm{Re}\{\mathrm{e}^{(\mathrm{i}(\Delta-\varphi))}\}\\&=\frac{A\omega^2}{R}\cos(\Delta-\varphi).\end{aligned}$$

(8)

The maximum amplitude of the record, U, is when $\cos(\Delta-\varphi)=1$. The corresponding phase lag Δ between the recorded signal and the ground motion is

$$\Delta=\tan^{-1}\left(\frac{2\lambda\omega\omega_0}{\omega_0^2-\omega^2}\right),$$

(9)

and the equation for the amplitude of the seismic record is given by

$$u=\frac{A\omega^2}{[(\omega_0^2-\omega^2)^2+4\lambda^2\omega^2\omega_0^2]^{1/2}}\cos(\omega t-\Delta).$$

(10)

6.5.3.1 Effect of Instrumental Damping

The ground motion caused by a seismic wave contains a broad spectrum of frequencies. Equation (6.111) shows that the response of the seismometer to different signal frequencies is strongly dependent on the value of the damping factor λ (Fig. 6.48). A completely *undamped* seismometer has $\lambda = 0$, and for small values of λ the response of the seismometer is said to be *underdamped.* An undamped seismometer preferentially amplifies signals near the natural frequency, and therefore cannot make an accurate record of the ground motion; the undamped instrument will resonate at its natural frequency ω_0. For all damping factors $\lambda < 1/\sqrt{2}$ the instrument response function has a peak, indicating preferential amplification of a particular frequency.

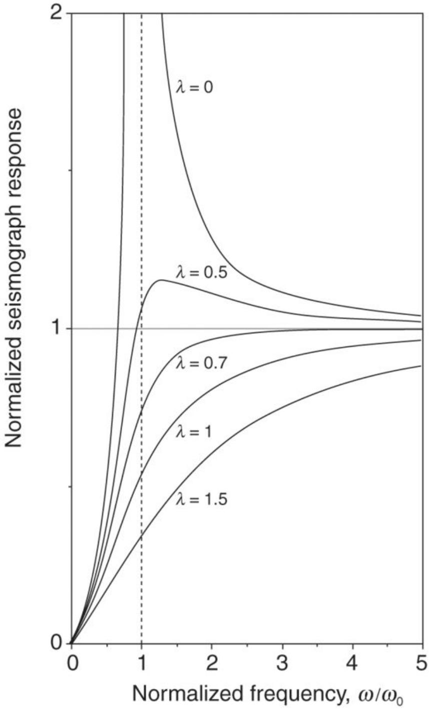


Fig. 6.48 Effect of the damping factor λ on the response of a seismometer to different signal frequencies. Critical damping corresponds to $\lambda = 1$. Satisfactory operation corresponds to a damping factor between 0.7 and 1 (i.e., 70–100% of critical damping).

The value $\lambda = 1$ corresponds to *critical damping*, so-called because it delineates two different types of seismometer response in the absence of a forcing vibration. If $\lambda < 1$, the damped, free seismometer responds to a disturbance by swinging *periodically* with decreasing amplitude about its rest position. If $\lambda \geq 1$, the disturbed seismometer behaves *aperiodically*, moving smoothly back to its rest position. However, if the damping is too severe ($\lambda \gg 1$), the instrument is *overdamped* and all frequencies in the ground motion are suppressed.

The optimum behavior of a seismometer requires that the instrument should respond to a wide range of frequencies in the ground motion, without preferential amplification or excessive suppression of frequencies. This requires that the damping factor should be close to the critical value. It is usually chosen to be in the range 70–100% of critical damping (i.e., $1/\sqrt{2} \leq \lambda < 1$). At critical damping the response of the seismometer to a periodic disturbing signal with frequency ω is given by

$$u = \frac{A\omega^2}{(\omega_0^2 + \omega^2)} \cos(\omega t - \Delta). \tag{6.112}$$

6.5.3.2 Long-Period and Short-Period Seismometers

The natural period ($2\pi/\omega_0$) of a seismometer is an important factor in determining what it actually records. Two examples of special interest correspond to instruments with very long and very short natural periods, respectively.

The long-period seismometer is an instrument in which the resonant frequency ω_0 is very low. For all but the lowest frequencies we can write $\omega_0 \ll \omega$. The phase lag Δ between the seismometer and the ground motion becomes zero (see Box 6.2), and the amplitude of the seismometer displacement becomes equal to the amplified ground displacement q:

$$u = A \cos\omega t = q. \tag{6.113}$$

The long-period seismometer is sometimes called a *displacement meter*. It is usually designed to record seismic signals with frequencies of 0.01–0.1 Hz (i.e., periods in the range 10–100 s).

The short-period seismometer is constructed so as to have a very short natural period and a correspondingly high resonant frequency ω_0, which is higher than most frequencies in the seismic wave. Under these conditions we have $\omega_0 \gg \omega$, the phase difference Δ is again small, and Eq. (6.112) becomes

$$u = \frac{\omega^2}{\omega_0^2} A \cos\omega t = -\frac{1}{\omega^2}\ddot{q}. \tag{6.114}$$

This equation shows that the displacement of the short-period seismometer is proportional to the acceleration of the ground, and the instrument is accordingly called an *accelerometer*. It is usually designed to respond to seismic frequencies of 1–10 Hz (periods in the range 0.1–1 s). An accelerometer is particularly suitable for recording strong-motion earthquakes, when the amplitude of the ground motion would send a normal type of displacement seismometer off-scale.

6.5.3.3 Broadband Seismometers

Short-period seismometers operate with periods of 0.1–1 s and long-period instruments are designed for periods greater than 10 s, omitting a period range that is interesting, for instance, in studies of regional crustal structure. This handicap was overcome by the design of broadband seismometers.

The broadband seismometer has basically an inertial pendulum-type design, with enhanced capability due to a force-feedback system. This works by applying a force proportional to the displacement of the inertial mass to prevent it from moving significantly. The amount of feedback force applied is determined by using an electrical transducer to convert motion of the mass into an electrical signal. The force needed to hold the mass stationary corresponds to the ground acceleration. The feedback electronics are critical to the success of this instrument.

Broadband design results in a seismometer with great bandwidth. The recording of an earthquake by a broadband seismometer contains more useable information than can be obtained from the short-period or long-period recordings individually or in combination (Fig. 6.49).

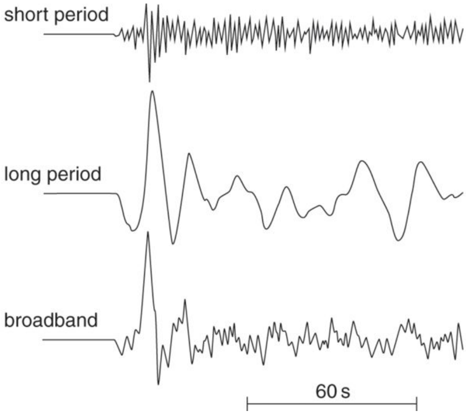


Fig. 6.49 Comparison of short-period and long-period records of a teleseismic P-wave with a broadband seismometer recording of the same event, which contains more information than the other two records separately or combined.

Source: T. Lay and T.C. Wallace, *Modern Global Seismology*, 515 pp., San Diego: Academic Press, 1995.

6.5.4 Data Acquisition

While the large majority of seismic studies employs inertial seismometers that operate under the same basic principles, data acquisition geometries and the types of instruments are very strongly application-dependent. Seismic exploration that aims to characterize reservoirs may use hundreds to hundreds of thousands of receiver locations, occupied by robust, low-cost, and very short-period seismometers called *geophones*. In contrast, seismological investigations of the deeper crust, mantle, and core often rely on smaller numbers of highly sophisticated broadband seismometers, designed to record the whole frequency spectrum generated by earthquakes.

6.5.4.1 Seismic Exploration

Seismic exploration largely relies on reflection seismology, i.e., the detection of reflectors and the estimation of seismic velocities of subsurface layers. The principle is simple. A seismic signal is produced by an explosion at a known place at a known time, and the reflections from the layer boundaries are recorded and analyzed. Geophones are positioned in the region of subcritical reflection, within the critical distance from the shot-point, where no refracted arrivals are possible. In addition to the wave that travels directly from the shot-point to the geophones and the waves reflected at subsurface interfaces, surface waves are also recorded. They may interfere with the reflected waves and are often suppressed by sophisticated data processing. Reflection seismic data are often acquired along profiles, but in recent years, two-dimensional surveys, which cover a large area, have become more important.

Several field procedures are in common use. They are distinguished by different layouts of the geophones relative to the shot-point. The most routine application of reflection seismology is in *continuous profiling*, in which the geophones are laid out at discrete distances along a profile through the shot-point. After each shot the geophone layout and shot-point are moved a predetermined distance along the profile, and the procedure is repeated. Broadly speaking, there are two main variations of this method, depending on whether each reflection point on the reflector is sampled only once (*conventional coverage*) or more than once (*redundant coverage*).

A common form of conventional coverage is a *split-spread* method (Fig. 6.50), in which the geophones are spread symmetrically on either side of the shot-point. If the reflector is flat-lying, the point of reflection of a ray recorded at any geophone is below the point midway between the shot-point and the geophone. For a shot-point at Q the rays QAP and QBR that are reflected to geophones at P and R represent extreme cases. The two-way travel time of the ray QAP gives the depth of the reflection point A, which is plotted below the mid-point of QP. Similarly, B is plotted below the mid-point of QR. The split–spread layout around the shot-point Q gives the depths of reflection points along AB, which is half the length of the geophone spread PR. The shot-point is now moved to the point R, and the geophones between P and Q are moved to cover the segment RS. From the new shot-point R the positions of reflection points in the segment BC of the reflector are obtained. The ray RBQ from shot-point R to the geophone at Q has the same path as the ray QBR from shot-point Q to the geophone at R. By successively moving the shot-point and half of the split-spread geophone layout, a continuous coverage of the subsurface reflector is obtained.

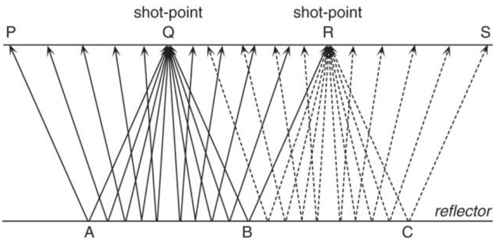


Fig. 6.50 The split–spread method of obtaining continuous

subsurface coverage of a seismic reflector.

Redundant coverage is illustrated by the *common-mid-point* method, which is routinely employed as a means of reducing noise and enhancing the signal-to-noise ratio. The principle is illustrated for a small number of 11 geophones in Fig. 6.51. When a shot is fired at A, the signals received at geophones 3–11 give subsurface coverage of the reflector between points a and e. The shot-point is now moved to B, which coincides with the position occupied by geophone 2 for the first shot, and the geophone array is moved forward correspondingly along the direction of the profile to positions 4–12. From shot-point B, the subsurface coverage of the reflector is between points b and f. The reflector points b to e are common to both sets of data. By repeatedly moving the shot-point and geophone array in the described manner, each reflecting point of the interface is sampled multiply. For example, in Fig. 6.51 the reflecting point d is sampled multiply by the rays Ad9, Bd8, Cd7, etc. The lengths of these ray paths are different. During subsequent data-processing the reflection travel times are corrected for *normal moveout*, which is a geometrical effect related to geophone distance from the shot-point. The records are then *stacked* (summed or averaged) in order to enhance the signal-to-noise ratio.

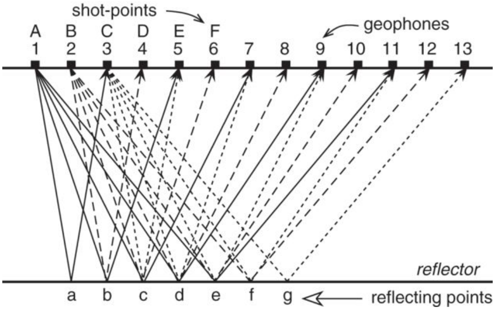


Fig. 6.51 Common-mid-point method of seismic reflection shooting, showing rays from successive shot-points at A, B, and C, and the repeated sampling of the same point on the reflector (e.g., d) by rays from each shot-point.

6.5.4.2 Regional and Global Seismology

The backbone of many seismological studies are broadband seismometers that operate over many years, are carefully insulated, cemented to the ground, and removed from sources of anthropogenic noise (if possible). Currently, the number of such *permanent stations* that have been operational continuously during the past 10 years is on the order of several thousand, worldwide.

The global network of permanent stations is complemented by arrays of *temporary stations*. A temporary array may be installed for a few weeks up to a few years in order to study the structure and seismic activity of a specific region. A special type of temporary array was developed in the early 1990s in order to cover entire continents. The *Skippy array*, named after the Australian bush kangaroo, was a group of around 10 seismometers successively installed in different parts of Australia between 1992 and 1996. “Hopping” from east to west, the whole continent was ultimately covered. The Skippy concept continues to be used. Fig. 6.52 shows more than 2000 locations successively occupied by around 400 broadband seismometers between 2007 and 2013, as part of the *USArray* initiative. The average receiver spacing was 70 km.

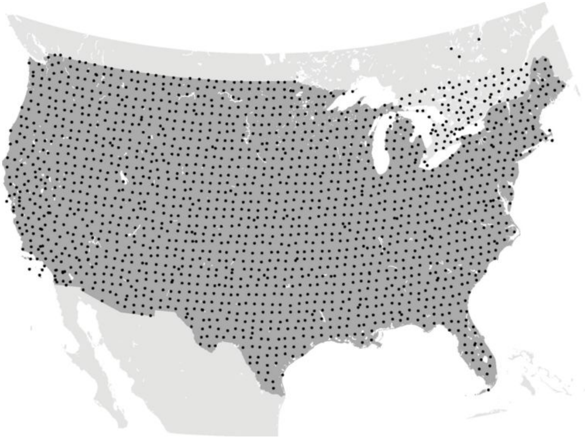


Fig. 6.52 Locations of USArray stations (black dots). The actual array consisted of around 400 broadband seismometers that were moved from west to east between 2007 and 2013.

6.5.5 Developments in Seismic Instrumentation

Despite the availability of conventional high-quality inertial seismometers, new seismic instrumentation continues to be developed. For example, ocean-bottom seismometers (OBS) are increasingly being used to improve coverage in the oceans. Typically, an OBS is dropped (gently) from a ship, sinks to the ocean floor, and remains there until its recovery, weeks or months later. In contrast to an on-land seismometer, an OBS cannot communicate during its deployment. The absence of GPS signals requires highly accurate clocks that do not drift significantly in response to temperature changes. Electricity must be provided by powerful batteries (instead of solar panels commonly used on land), and all recorded data must be stored on disk.

A recently developed alternative are seismometers installed in *floats* (Fig. 6.53). Floating with the ocean current at several hundred meter depths, the on-board seismometer records acoustic pressure variations. When a signal resembles a P-wave (and not the singing of a whale, for example), the float ascends to the surface where its current position is determined by GPS. After transmitting the data via satellite communication, the float dives back and continues monitoring.

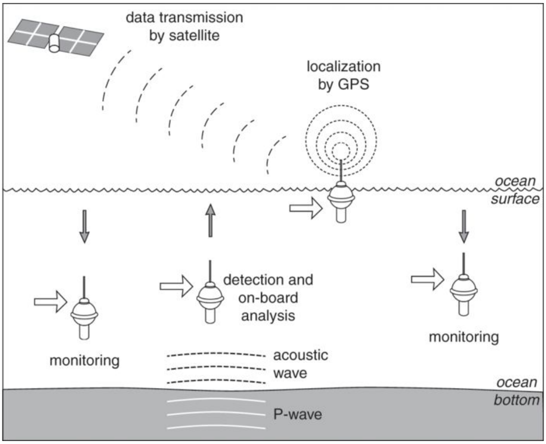


Fig. 6.53 Illustration of a float recording acoustic waves that have been generated by P-waves at the ocean bottom. Horizontal arrows indicate ocean currents (courtesy of Guust Nolet).

A new measurement technology known as *distributed acoustic sensing* (DAS) abandons the principle of inertial seismometers and records ground deformation optically (Fig. 6.54). An interrogation unit (IU) sends short laser pulses into an optical fiber. The fiber contains impurities which produce secondary waves, in accord with Huygens' principle (Section 6.3.2). The secondary waves, also called *scattered waves*, propagate back into the IU where they are recorded. When the fiber is stretched, for example in response to a seismic wave passing by, the impurities move slightly. Consequently, the scattered light acquires a small time or phase shift, also recorded by the IU. The measurement of phase shifts as a function of time can then be related to the strain ε along the fiber. While still in its infancy, DAS is regarded as a particularly promising future technology because telecommunication fibers that already exist may be used.

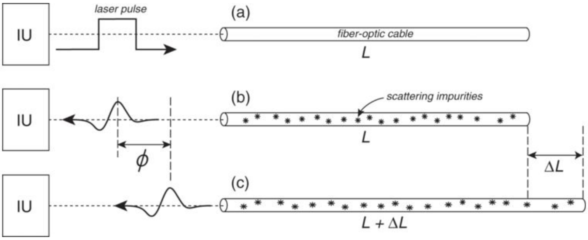


Fig. 6.54 Illustration of distributed acoustic sensing (DAS). IU is the interrogation unit, and Φ is the phase shift of the scattered light after the fiber has been stretched from length L to $L + \Delta L$ (courtesy of Patrick Paitz).

Laser optics can also be used to measure ground rotation, instead of displacement u or strain ε. For this, fibers are arranged in circular loops. When the loop rotates while the laser pulse is traveling through it, the distance to the IU changes. The resulting time shift may then be related to the absolute rotation of the device.

6.6 *Suggestions for Further Reading*

Intermediate and Advanced Level

Aki, K. and Richards, P. G. 2002. *Quantitative Seismology*. Sausalito, CA: University Science Books.

Bullen, K. E. and Bolt, B. A. 1985. *An Introduction to the Theory of Seismology*. Cambridge: Cambridge University Press.

Chapman, C. 2004. *Fundamentals of Seismic Wave Propagation*. Cambridge: Cambridge University Press.

Gubbins, D. 1990. *Seismology and Plate Tectonics*. Cambridge: Cambridge University Press.

Lay, T. and Wallace, T. C. 1995. *Modern Global Seismology*. San Diego, CA: Academic Press.

Shearer, P. 2019. *Introduction to Seismology* (3rd ed.). Cambridge: Cambridge University Press.

Stein, S. and Wysession, M. 2003. *An Introduction to Seismology, Earthquakes and Earth Structure*. Oxford: Blackwell Publishing.

Udias, A. 2000. *Principles of Seismology*. Cambridge: Cambridge University Press.

6.7 Review Questions

1. Describe the principle of the inertial seismometer.

2. Describe the particle motions relative to the direction of propagation of the two seismic body waves and the two seismic surface waves.

3. What are the *Lamé* constants in elasticity theory?

4. What is meant by *dispersion* with regard to the propagation of surface waves?

5. Sketch how Rayleigh waves from a distant earthquake would appear on a seismogram.

6. What are the free oscillations of the Earth? What kinds are possible?

7 What is meant by the normal *modes* of an oscillation? What are *higher modes*? How do they relate to surface waves?

8. What is the *critical distance* in seismic refraction surveying? What is the *crossover distance*? What is a *head wave*? What are supercritical reflections?

9. A seismic survey is conducted over level ground consisting of two horizontal layers. Sketch a travel time diagram that shows the arrivals of the direct wave, the reflected wave, and the doubly refracted wave. How can the seismic velocity of each layer be determined from the diagram?

10. What is *normal moveout* in seismic reflection profiling?

11. What is meant by the migration of reflection seismic records? Why is it necessary?

12. Describe the *split-spread* and *common-mid-point* methods in reflection profiling. What is achieved by each of these methods?

6.8 Exercises

1. Calculate the bulk modulus (K), the shear modulus (μ), and Poisson's ratio (ν) for the lower crust, upper mantle, and lower mantle, respectively, using the values for the P-wave (α) and S-wave (β) velocities, and density (ρ) in the following table.

Region	**Depth (km)**	**α (km s^{-1})**	**β (km s^{-1})**	**ρ (kg m^{-3})**
Lower crust	33	7.4	4.3	3100
Upper mantle	400	8.5	4.8	3900
Lower mantle	2200	12.2	7.0	5300

2. The table below gives the densities and seismic P- and S-wave velocities at various depths in the Earth.

Depth (km)	**ρ (1000 kg m^{-3})**	**α (km s^{-1})**	**β (km s^{-1})**
100	3.38	8.05	4.45
500	3.85	9.65	5.22
1000	4.58	11.46	6.38
2000	5.12	12.82	6.92
2890	5.56	13.72	7.27
2900	9.90	8.07	0
4000	11.32	9.51	0
5000	12.12	10.30	0
5500	12.92	11.14	3.58
6470	13.09	11.26	3.67

(a) From these quantities, calculate the rigidity modulus (μ), bulk modulus (K), and Poisson's ratio (ν) at each depth.

(b) Discuss in your own words the information that these data give about the deep interior of the Earth.

3. Why might one expect an interface with a small critical angle to be a good reflector of seismic energy?

4. A plane seismic wave, traveling vertically downwards in a rock of density 2200 kg m^{-3} with seismic velocity 2000 m s^{-1}, is incident on the horizontal top surface of a rock layer of density 2400 kg m^{-3} and seismic velocity 3300 m s^{-1}.

(a) What are the amplitude ratios of the transmitted and reflected waves?

(b) What fraction of the energy of the incident wave is transmitted into the lower medium?

5. A plane seismic wave travels vertically downwards at a velocity of 4800 m s^{-1} through a salt layer with density 2100 kg m^{-3}. The wave is incident upon the top surface of a sandstone layer with density 2400 kg m^{-3} The phase of the reflected wave is changed by 180° and the reflected amplitude is 2% of the incident amplitude. What is the seismic velocity of the sandstone?

6. (a) Calculate the minimum arrival times for seismic reflections from each of the reflecting interfaces in the following section. Consider the base of the lowermost bed to be also a reflector.

Formation	Density (kg m^{-3})	Thickness (m)	Formation velocity (m s^{-1})
Alluvium	1500	150	600
Shale	2400	450	2700
Sandstone	2300	600	3000
Limestone	2500	900	5400
Salt	2200	300	4500
Dolomite	2700	600	6000

(b) What is the average velocity of the section for a reflection from the base of the dolomite?

(c) Using the listed densities, calculate the reflection coefficient for each interface (except the base of the dolomite). Which interface gives the strongest reflection and which the weakest? At which interfaces does a change in phase occur? What does this mean?

7. A reflection seismic record in an area of relatively flat dips gave the following data:

Distance shot-point to detector (m)	**Travel-time (t_1), first reflection (s)**	**Travel time (t_2), second reflection (s)**
30	1.000	1.200
90	1.002	1.201
150	1.003	1.201
210	1.007	1.202
270	1.011	1.203
330	1.017	1.205
390	1.023	1.207

(a) Plot the t–x curves for these reflections to show the "moveout" effect.

(b) On a different graph, plot the t^2–x^2 curves (i.e., squared data) for the reflections.

(c) Determine the average vertical velocity from the surface to each reflecting bed.

(d) Use these velocities to compute the depths to the reflecting beds.

8. The following table gives two-way travel times of seismic waves reflected from different reflecting interfaces in a horizontally layered medium.

Geophone to shot-point distance (m)	**Travel time (s) to**		
first reflector	**second reflector**	**third reflector**	
500	0.299	0.364	0.592
1000	0.566	0.517	0.638
1500	0.841	0.701	0.708

2000	1.117	0.897	0.799
2500	1.393	1.099	0.896

(a) Draw a plot of (travel time)^2 against (distance)^2.

(b) Determine the vertical two-way travel time ("echo time") and average velocity to each reflecting interface.

(c) Compute the depth of each reflector and the thickness of each layer.

(d) Compute the true velocity (interval velocity) of each layer.

(e) Verify your results by computing the total vertical travel time for a wave reflected from the deepest interface.

9. Assume the horizontally layered structure from the previous problem.

(a) If a seismic ray leaves the surface at an angle of 15° to the vertical, how long does it take to return to the surface after reflecting from the basement?

(b) At what horizontal distance from the shot-point does this ray reach the surface?

10. Assume that the three horizontal homogeneous rock layers in the previous problems have densities of 1800, 2200, and 2500 kg m^{-3}, respectively. The lowest layer overlies basement with velocity 5.8 km s^{-1} and density 2700 kg m^{-3}.

(a) Compute the reflection and transmission coefficients at each interface for a plane P-wave traveling vertically downwards.

(b) Calculate what fraction of the initial energy of the wave is transmitted into the basement.

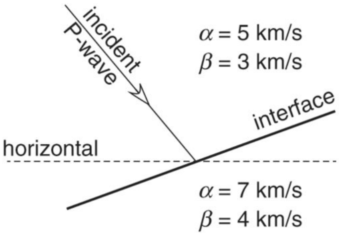
(c) Calculate the fraction of the initial energy carried in the reflection that returns to the surface from the basement.

11. An incident P-wave is converted into refracted and reflected

P- and S-waves at an interface. Calculate all the critical angles in the following three cases, where α and β are the P-wave and S-wave velocities, respectively:

Layer	Seismic wave	Case (a) (km s^{-1})	Case (b) (km s^{-1})	Case (c) (km s^{-1})
Above interface	α	3.5	4.0	5.5
β	2.0	2.3	3.1	
Below interface	α	8.5	6.0	7.0
β	5.0	3.5	4.0	

12. An incident P-wave is converted into refracted and reflected P- and S-waves at an interface that is inclined at 20° to the horizontal, as in the figure below. The respective P- and S-wave velocities are 5 km s^{-1} and 3 km s^{-1} above the interface and 7 km s^{-1} and 4 km s^{-1} below the interface. If the incident P-wave strikes the interface at an angle of 40° to the *horizontal*, calculate the angles to the *horizontal* made by the reflected and refracted P- and S-waves.



13. A seismic refraction survey gave the following data for the first arrival times at various distances from the shot-point.

Distance (km)	Time (s)	Distance (km)	Time (s)
3.1	1.912	13.1	6.678
5.0	3.043	14.8	7.060
6.5	3.948	16.4	7.442
8.0	4.921	18.0	7.830
9.9	5.908	19.7	8.212

11.5	6.288

(a) Plot the travel time curve for the first arrivals.

(b) Calculate the seismic velocities of the layers.

(c) Calculate the critical angle of refraction for the interface.

(d) Calculate the minimum depth to the refracting interface.

(e) Calculate the critical distance for the first arrival of refracted rays.

(f) Calculate the crossover distance beyond which the first arrivals correspond to head waves.

14. A seismic refraction survey is carried out over a layered crust with flat-lying interfaces. In one case the crust is homogeneous and 30 km thick with a P-wave velocity 6 km s$_{–}^{1}$, and overlies mantle with P-wave velocity 8 km s$_{–}^{1}$. In the other case the crust consists of an upper layer 20 km thick with P-wave velocity 6 km s$_{–}^{1}$ overlying a lower layer 10 km thick with P-wave velocity 5 km s$_{–}^{1}$. The upper mantle P-wave velocity is again 8 km s$_{–}^{1}$. Plot the first arrival time curves for the two cases on the same graph. What is the effect of the low-velocity layer on the estimation of depth to the top of the mantle?

15. The table below gives up-dip and down-dip travel times of P-wave arrivals for refraction profiles over an inclined interface. The geophones are laid out in a straight line passing through the alternate shot-points A and B, which are 2700 m apart on the profile.

Distance from shot-point (m)	**Travel time (s)**	
from A	**from B**	
300	0.139	0.139
600	0.278	0.278
900	0.417	0.417

1200	0.556	0.556
1500	0.695	0.695
1800	0.833	0.833
2100	0.972	0.972
2400	1.085	1.111
2700	1.170	1.170
3000	1.255	1.223
3300	1.339	1.276
3600	1.424	1.329

(a) Plot the travel time curves for each shot-point.

(b) Calculate the true velocity of the upper layer.

(c) Calculate the apparent velocities of the layer below the refractor.

(d) In which direction does the refracting interface dip?

(e) What is the angle of dip of the interface?

(f) What is the true velocity of the layer below the refractor?

(g) What are the closest distances to the refractor below A and B?

(h) What are the vertical depths to the refractor below A and B?

6.9 Computational Exercises

Jupyter notebook computational exercises can be found online for this chapter at www.cambridge.org/FoG3ed ((1) Reflections at dipping interface, (2) Finite-difference solution of the wave equation).

7

Earthquakes and the Earth's Internal Structure

Preview

Large earthquakes are among the most energetic phenomena observed on Earth, surpassing the biggest nuclear explosions by orders of magnitude. The family of earthquakes is very diverse. It includes natural events that occur within seconds along the boundaries of tectonic plates, slow earthquakes that last for days or weeks, swarms of thousands of little earthquakes beneath volcanoes, and clusters of earthquakes induced by industrial activity. Though individual earthquakes cannot be predicted with current methods, long-term seismic hazard can be estimated in order to inform building codes and to prepare the population. Seismic waves emitted by earthquakes travel through the Earth, thereby acquiring information about its internal structure. Seismic tomography based on earthquake recordings draws the image of a very dynamic planet, featuring cold lithospheric slabs that descend deep into the mantle, and narrow plumes that transport hot material toward the surface.

7.1 Introduction

Early descriptions of earthquake characteristics were necessarily restricted to observations and measurements in the “near-field” region of the earthquake, i.e., in comparatively close proximity to the place where it occurred. A conspicuous advance in the science of seismology was accomplished with the invention of a sensitive and reliable seismometer by John Milne in 1892. Although massive and primitive by comparison with modern instruments, the precision and sensitivity of this revolutionary new device permitted accurate, quantitative descriptions of earthquakes at large distances from their source, in their “far-field” region. The accumulation of reliable records of distant earthquakes (designated as *teleseismic* events) made possible the systematic study of the Earth’s seismicity and its internal structure.

The great San Francisco earthquake of 1906 was intensively studied and provided an impetus to efforts at understanding the origin of these natural phenomena, which were partly explained in the same year by the elastic rebound model of Harry F. Reid. Also, in 1906, Richard Dixon Oldham proposed that the best explanation for the travel times of teleseismic waves through the body of the Earth required a large, dense, and probably fluid core; the depth to its outer boundary was calculated in 1913 by Beno Gutenberg. From the analysis of the travel times of seismic body waves from near earthquakes in former Yugoslavia, Andrija Mohorovičić in 1909 inferred the existence of the crust–mantle boundary, and in 1936 the existence of the solid inner core was deduced by Inge Lehmann. The definitions of these and other discontinuities associated with the deep internal structure of the Earth have since been greatly refined.

The needs of the world powers to detect incontrovertibly the testing of nuclear bombs by their adversaries provided considerable stimulus to the science of seismology in the 1950s and 1960s. The amount of energy released in a nuclear explosion can be comparable to that of an earthquake, but the phenomena can be discriminated by analyzing the directions of first motion recorded by seismometers. The accurate location of the event required improved knowledge of seismic body-wave velocities throughout the Earth's interior. The political necessities of the Cold War led to major improvements in seismological instrumentation, and to the establishment of a new worldwide network of seismic stations with the same physical characteristics. These developments had an important feedback to the earth sciences, because they resulted in more accurate locations of earthquake epicenters and a better understanding of the Earth's structure. The pattern of global seismicity, with its predominant concentration in narrow active zones, was an important factor in the development of the theory of plate tectonics, as it allowed the identification of plate margins and the sense of relative plate motions.

7.2 Earthquake Seismology

7.2.1 Types of Earthquakes

By far the largest and most numerous earthquakes originate from the sudden movement of faults separating tectonic plates. In addition to such *tectonic earthquakes*, measurable ground motion may be caused by landslides, the collapse of underground caverns and volcanic calderas, volcanic and man-made explosions, or meteoroid bursts in the atmosphere. The latter can be observed around three times per day.

Short-lived tectonic earthquakes that typically do not last for more than a few seconds are contrasted by *volcanic tremor* with durations up to several minutes. Volcanic tremor is often observed at very specific frequencies, giving it the appearance of a long-lasting sine curve. Its most likely origin is the rapid movement of gases and fluids within the magmatic system. Frequently, volcanic tremor accompanies or precedes volcanic eruptions, which makes it a valuable forecasting tool.

The emergence of highly sensitive seismometer networks revealed the existence of a different type of tremor, unknown prior to 1999 (more than 100 years after Milne's seismometer). *Nonvolcanic* or *tectonic tremor* is associated with large faults that are capable of producing major earthquakes. While the seismic signal of nonvolcanic tremor may appear similar to volcanic tremor of long duration, it has been shown to consist of a large number of individual small earthquakes, with magnitudes around 3 or lower (see Section 7.2.6.2 for a precise definition of earthquake magnitude). In contrast to ordinary tectonic earthquakes, these swarm events hardly generate waves with frequencies above 3 Hz, which is why they are commonly referred to as *low-frequency earthquakes*.

Around the same time that nonvolcanic tremor was discovered, scientists started to observe distinct episodes of displacement on GPS stations located close to subduction zones. One of these episodes in 1999 involved a 50 × 300 km patch of the Cascadia subduction zone off the coast of western North America. It lasted for 35 days, producing a total displacement of around 2 cm. Owing to the slow occurrence of slip, compared to ordinary tectonic earthquakes, these events are called *slow-slip events*. Since slow-slip events hardly generate seismic waves that can be recorded by seismometers, they are often said to be *quiet* or even *silent earthquakes*. Following the initial discovery in Cascadia, slow-slip events have meanwhile been identified at other tectonic plate boundaries, e.g., in Japan, New Zealand, Mexico, and Alaska. Slow-slip events show a remarkable, almost clock-like, recurrence pattern. The events in Cascadia repeat every 13–16 months, with an average of 14.5 months. They generally last 2–4 weeks. Slow-slip events in the Nankai subduction zone in southwest Japan repeat approximately every six months.

Though low-frequency earthquakes and slow-slip events were initially thought to be independent phenomena, it is now widely accepted that they are closely related. Slow-slip events tend to be accompanied by clusters of low-frequency earthquakes that we perceive as nonvolcanic tremor. The locus of slow-slip events can migrate several kilometers per day, and the associated low-frequency earthquakes migrate with them.

7.2.2 Elastic Rebound Model

Most earthquakes are so weak that they are only registered by sensitive seismometers, but some are strong enough to have serious, even catastrophic, consequences for mankind and the environment. The majority of all earthquakes result from tectonic events, primarily movements on faults. The remaining events are related to volcanism, collapse of subterranean cavities, or man-made effects.

Our understanding of the processes that lead to earthquakes derives to a large extent from observations of seismic events on large strike–slip faults that are visibly exposed at the surface, such as the San Andreas fault in California and the North Anatolian fault in Turkey. The average relative motion of the plates adjacent to the San Andreas fault is about 5 cm/yr, with the block to the west of the fault moving northward. On the fault plane itself, this motion is not continuous but takes place spasmodically. According to plate tectonic theory this extensively studied fault system is a transform fault. This is a rather special type, so it cannot be assumed that the observations related to the San Andreas fault are applicable without reservation to all other faults. However, the *elastic rebound model* proposed by Harry F. Reid after the 1906 San Francisco quake is a useful guide to how an earthquake may occur.

The model is illustrated in Fig. 7.1 by the changes to five parallel lines, drawn normal to the trace of the fault in the unstrained state and intersecting it at points A–E. Strain due to relative motion of the blocks adjacent to the fault accumulates over several years. Far from the trace of the fault the five lines remain straight and parallel, but close to it they are bent. When the yield stress of the crustal rocks at C is exceeded (Section 5.3.4), brittle failure occurs and there is a violent rupture on the fault plane. The relative displacement that has been taking place progressively between the adjacent plates during years or decades is achieved on the fault plane in a few seconds or minutes. The strained rocks adjacent to the fault “rebound” suddenly. The accumulated strain energy is released with the seismic velocity of the ruptured rocks, which is several km s^{-1}. The segments BC and C′D undergo compression, while CD and BC′ experience dilatation. The points A and E do not move; the stored strain energy at these points is not released. The entire length of the fault plane is not displaced, only the region in which the breaking point has been exceeded. The greater the length of the fault plane that is activated, the larger is the ensuing earthquake.

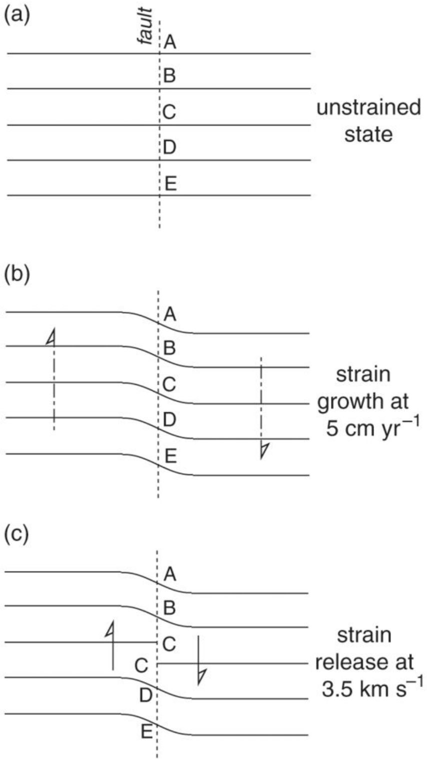

(a)
fault
A
B
C
D
E
unstrained state
(b)
A
B
C
D
E
strain growth at 5 cm yr^{-1}
(c)
A
B
C
C
D
E
strain release at 3.5 km s^{-1}

Fig. 7.1 Elastic rebound model of the origin of earthquakes: (a) unstrained state of a fault segment; (b) accumulation of strain close to the fault due to relative motion of adjacent crustal blocks; and (c) “rebound” of the strained segment as an earthquake with accompanying release of seismic energy.

The December 26, 2004 Sumatra–Andaman earthquake ruptured a fault exceeding 1200 km in length, with a displacement across the fault reaching 20 m in some places. The vast majority of earthquakes are fortunately much smaller, involving a fault plane measuring several square kilometers in area. From the point of view of an observer at a distance of hundreds or even thousands of kilometers such earthquakes appear to happen at a point. This point is called the *focus* or *hypocenter* of the earthquake (Fig. 7.2). It generally occurs at a *focal depth* many kilometers below the Earth's surface. The point on the Earth's surface vertically above the focus is called the *epicenter* of the earthquake.

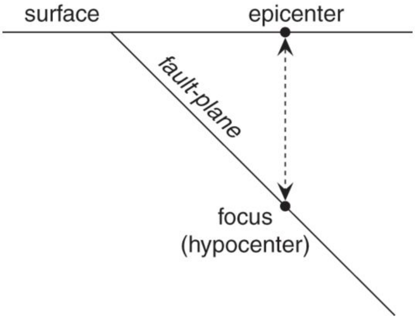


Fig. 7.2 Vertical section perpendicular to the plane of a normal fault, defining the epicenter and hypocenter (focus) of an earthquake.

The occurrence of a large earthquake is not necessarily as abrupt as described in the preceding paragraph, although it can be very sudden. In 1976 a major earthquake with magnitude 7.8 struck a heavily populated area of northern China near the city of Tangshan. Although there were known faults in the area, they had long been seismically inactive, and the large earthquake struck without warning. It completely devastated the industrial region and caused an estimated 243,000 fatalities. However, in many instances the accumulating strain is partially released locally as smaller earthquakes, which may indicate that strain energy is building up to a level that permits the occurrence of larger events. Smaller earthquakes that immediately precede a larger one are often referred to as *foreshocks*, though a causal relation between them is often hard to establish.

When an earthquake occurs, most of the stored energy is released in the main shock. The rapidly changing deformation generally triggers large numbers of smaller earthquakes, known as *aftershocks*, which may last for weeks or months. Some aftershocks can be comparable in size to the main earthquake. The death toll from aftershocks is likely to be less, because people have evacuated damaged structures.

7.2.3 Location of the Epicenter of an Earthquake

The distance of a seismic station from the epicenter of an earthquake (the *epicentral distance*) may be expressed in kilometers Δ_{km} along the surface, or by the angle $\Delta^{\circ} = (180/\pi)(\Delta_{\mathrm{km}}/R)$ subtended at the Earth's center. The travel times of P- and S-waves from an earthquake through the body of the Earth to an observer are dependent on the epicentral distance (Fig. 7.3a). The travel time versus distance plots are not linear, because ray paths of waves traveling to distant seismometers are curved. However, the seismic velocity profile of the Earth's interior is known well enough that the travel times for each kind of wave can be tabulated or graphed as a function of epicentral distance. In computing epicentral distance from earthquake data, the total travel time is not at first known, because an observer is rarely at the epicenter to record the exact time of occurrence t_0 of the earthquake. However, the difference in travel times for P- and S-waves $(t_{\mathrm{s}} - t_{\mathrm{p}})$ can be obtained directly from the seismogram; it increases with increasing epicentral distance (Fig. 7.3a).

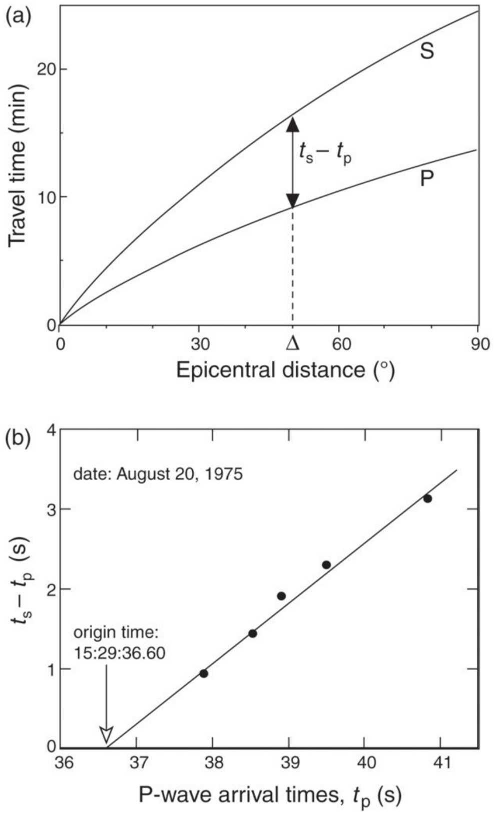


Fig. 7.3 (a) Travel times of P- and S-waves from an earthquake through the body of the Earth to an observer at epicentral distances up to 90°. The epicentral distance (Δ) of the earthquake is found from the difference in travel times ($t_s - t_p$). (b) Wadati diagram for determining the time of occurrence of an earthquake.

For local earthquakes we can assume that the seismic velocities α and β are fairly constant in the near-surface layers. The time when the earthquake occurred, t_0, can then be obtained by plotting the differences ($t_s - t_p$) against the arrival times t_p of the P-wave at different stations. The plot, called a *Wadati diagram*, is a straight line (Fig. 7.3b). If D is the distance traveled by the seismic wave, the travel times of P- and S-waves are respectively $t_p = D/\alpha$ and $t_s = D/\beta$, so

$$t_s - t_p = D\left(\frac{1}{\beta} - \frac{1}{\alpha}\right) = t_p\left(\frac{\alpha}{\beta} - 1\right). \tag{7.1}$$

The intercept t_0 of the straight line with the arrival-time axis is the time of occurrence of the earthquake and the slope of the line is $[(\alpha/\beta) - 1]$. Knowing the P-wave velocity α, the distance to the earthquake is obtained from $D = \alpha(t_p - t_0)$.

In order to determine the location of an earthquake, epicenter travel times of P- and S-waves to at least three seismic stations are necessary (Fig. 7.4). The data from one station give only the distance of the epicenter from that station. It could lie anywhere on a circle centered at the station. The data from an additional station define a second circle which intersects the first circle at two points, each of which could be the true epicenter. Data from a third station remove the ambiguity: the common point of intersection of the three circles is the epicenter.

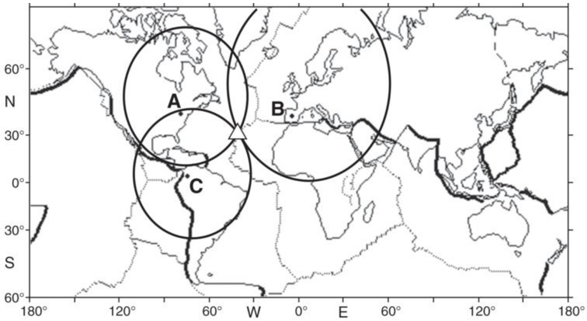


Fig. 7.4 Location of an earthquake epicenter using epicentral distances of three seismic stations (at A, B, and C). The epicentral distance of each station defines the radius of a circle centered on the station. The epicenter (triangle) is located at the common intersection of the circles, which appear oval due to the map projection.

Generally the circles do not intersect at a point, but form a small spherical triangle. The optimum location of the epicenter is at the center of the triangle. If data are available from more than three seismic stations, the epicentral location is improved; the triangle is replaced by a small polygon. This situation arises in part from observational errors, and because the theoretical travel times are imperfectly known. The interior of the Earth is neither homogeneous nor isotropic, as assumed above. The exact location of earthquake epicenters requires detailed knowledge of the seismic velocities along the entire path, but especially under the source area and the receiving station. The main reason for the intersection triangle or polygon is, however, that the seismic rays travel to the seismograph from the focus, and not from the epicenter. The focal depth of the earthquake, d, which may be up to several hundred kilometers, must be taken into account. It can be estimated from simple geometry. If Δ_{km} is the epicentral distance and D the distance traveled by the wave, to a first approximation $d = (D^2 - \Delta_{\mathrm{km}}{}^2)^{1/2}$. Combining several values of d from different recording stations gives a reasonable estimate of the focal depth.

7.2.4 Global Seismicity

The epicenters of around 600,000 earthquakes are now reported annually by the International Seismological Center (ISC). The geographical distribution of world seismicity, based on around 1.3 million earthquakes (Fig. 7.5), dramatically illustrates the tectonically active regions of the Earth. The seismicity map is important evidence in support of plate tectonic theory, and delineates the presently active plate margins.

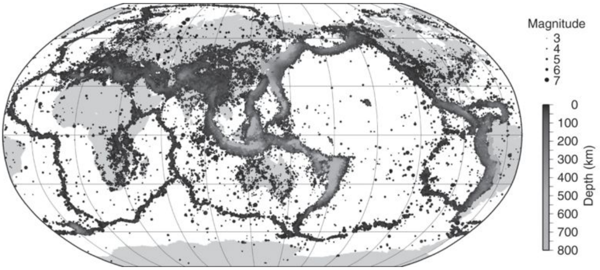


Fig. 7.5 World seismicity. Each of the 1.3 million dots on this map marks the epicenter of an earthquake located by the International Seismological Center (ISC). Grayscale indicates the depth of an earthquake. Note the increasing depth of subduction zone earthquakes, e.g., around the Pacific, which outline dipping oceanic lithosphere that sinks into the mantle of the Earth. (Courtesy of Kostas Lentas and James Harris, ISC.)

Earthquake epicenters are not uniformly distributed over the Earth's surface, but occur predominantly along rather narrow zones of *interplate* seismic activity. The *circum-Pacific zone*, in which about 75–80% of the annual release of seismic energy takes place, forms a girdle that encompasses the mountain ranges on the west coast of the Americas and the island arcs along the east coast of Asia and Australasia. The *Mediterranean–transasiatic zone*, responsible for about 15–20% of the annual seismic energy release, begins at the Azores triple junction in the Atlantic Ocean and extends along the Azores–Gibraltar ridge; after passing through North Africa it makes a loop through the Italian peninsula, the Alps, and the Dinarides; it then runs through Turkey, Iran, the Himalayan mountain chain, and the island arcs of Southeast Asia, where it terminates at the circum-Pacific zone. The system of *oceanic ridges and rises* form the third most active zone of seismicity, with about 3–7% of the annually released seismic energy. In addition to their seismicity, each of these zones is also characterized by active volcanism.

The remainder of the Earth is considered to be relatively *aseismic*. However, no region of the Earth can be regarded as completely earthquake-free. About 1% of the global seismicity is due to *intraplate* earthquakes, which occur remote from the major seismic zones. These are not necessarily insignificant: some very large and damaging earthquakes (e.g., the New Madrid, Missouri, earthquakes of 1811 and 1812 in the Mississippi river valley, with magnitudes exceeding 7.5) have been of the intraplate variety.

Earthquakes can also be classified according to their focal depths. Earthquakes with *shallow* focal depths less than 70 km occur in all the seismically active zones; only shallow earthquakes occur on the oceanic ridge systems. The largest proportion (about 85%) of the annual release of seismic energy is liberated in shallow-focus earthquakes. The remainder is set free by earthquakes with *intermediate* focal depths of 70–300 km (about 12%) and by earthquakes with *deep* focal depths greater than 300 km (about 3%). These occur mostly in the circum-Pacific and Mediterranean–transasiatic seismic zones, and accompany the process of plate subduction.

The distributions of epicentral locations and focal depths of intermediate and deep earthquakes give important evidence for the processes at a subduction zone. When the earthquake foci along a subduction zone are projected onto a cross-section normal to the strike of the plate margin, they are seen to define a zone of seismicity about 30–40 km thick in the upper part of the 80–100 km thick subducting oceanic plate, which plunges at roughly 30–60° beneath the overriding plate (Fig. 7.6). For many years the inclined seismic zone was referred to in Western literature as a *Benioff zone* in recognition of the Californian scientist, Hugo Benioff. In the years following World War II Benioff carried out important pioneering studies that described the distribution of deep earthquakes on steeply dipping surfaces of seismicity. Many characteristics of the occurrence of deep earthquakes had been described in the late 1920s by a Japanese seismologist, Kiyoo Wadati. He discovered that the closer the epicenters of earthquakes lay to the Asian continent, the greater were their focal depths; the deep seismicity appeared to lie on an inclined plane. It was Benioff, however, who in 1954 proposed as an explanation of the phenomenon that the ocean floor was being “subducted’ underneath the adjacent land. This was a bold proposal well in advance of the advent of plate tectonic theory. Today the zone of active seismicity is called a *Wadati–Benioff zone* in recognition of both discoverers.

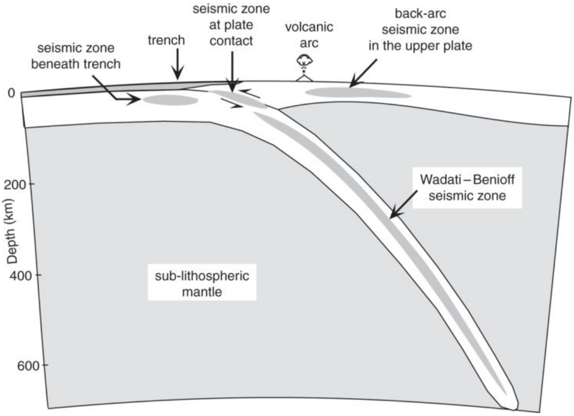

seismic zone at plate contact
back-arc seismic zone in the upper plate
volcanic arc
trench
seismic zone beneath trench
0
200
400
600
Depth (km)
Wadati–Benioff seismic zone
sub-lithospheric mantle

Fig. 7.6 Schematic cross-section through a subduction zone. The most active region is the zone of contact between the converging plates at depths of 10–60 km. There may be a “back-arc” seismic zone in the overriding plate. Below about 70 km depth a Wadati–Benioff seismic zone is described within the subducting plat.

Source: B.L. Isacks, Seismicity and plate tectonics, in: *The Encyclopedia of Solid Earth Geophysics*, D.E. James, ed., pp. 1061–1071, New York: Van Nostrand Reinhold, 1989.

In three dimensions the Wadati–Benioff zone constitutes an inclined slab dipping underneath the overriding plate. It marks the location and orientation of the upper surface of the subducting plate. The dip-angle of the zone varies between about 30° and 60°, becoming steeper with increasing depth, and it can extend to depths of several hundred kilometers into the Earth. The deepest reliably located focal depths extend down to about 680 km. Important changes in the crystalline structure of mantle minerals occur below this depth.

The structure of a subducting plate is not always as simple as described. A detailed study of the subducting Pacific plate revealed a double Wadati–Benioff zone under northeast Honshu, Japan (Fig. 7.7). The seismicity at depths below 100 km defines two parallel planes about 30–40 km apart. The upper plane, identified with the top of the subducting plate, is in a state of compression; the lower plane, in the middle of the slab, is in a state of extension. These stress states are the result of unbending of the subducting plate, which had previously undergone sharp bending at shallow depth below the trench axis. This information is inferred from analysis of the mechanisms by which the earthquakes occur.

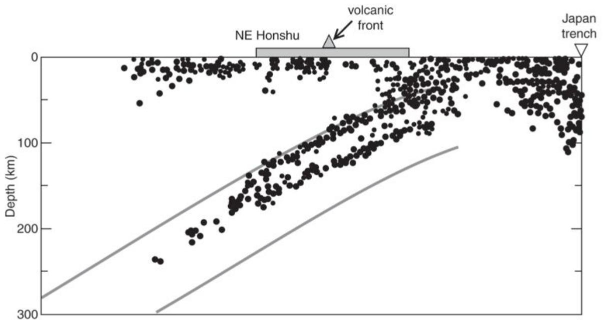


Fig. 7.7 Distribution of earthquakes in a vertical section across a double subduction zone under the island of Honshu, Japan. Below 100 km the seismicity defines two parallel planes, one at the top of the subducting plate, the other in the middle. The upper plane is in a state of compression, the lower is under extension.

Source: A. Hasegawa, Seismicity: subduction zone, in: *The*

Encyclopedia of Solid Earth Geophysics, D.E. James, ed., pp. 1054–1061, New York: Van Nostrand Reinhold, 1989.

7.2.5 Analysis of Earthquake Focal Mechanisms

During an earthquake the accumulated elastic energy is released suddenly by physical displacement of the ground, as heat, and as seismic waves that travel outward from the focus. By studying the first motions recorded by seismometers at distant seismic stations, the focal mechanism of the earthquake can be inferred and the motion on the fault plane interpreted.

Consider a vertical section perpendicular to the plane of a normal fault on which the hypocenter of an earthquake is located at point H (Fig. 7.8). When the region above the fault moves upslope, it produces a region of compression ahead of it and a region of dilatation (or expansion) behind it. In conjunction with the compensatory downslope motion of the lower block, the earthquake produces two regions of compression and two regions of dilatation surrounding the hypocenter. These are separated by the fault plane itself, and by an auxiliary plane through the focus and normal to the fault plane. When a seismic P-wave traveling out from a region of compression reaches an observer at C, its first effect is to *push* the Earth's surface upward; the initial effect of a P-wave that travels out from a region of dilatation to an observer at D is to *tug* the surface downwards. The P-wave is the earliest seismic wave to reach a receiver at C or D, and therefore the initial motion recorded by the instrument allows us to distinguish whether the first arrival was compressional or dilatational.

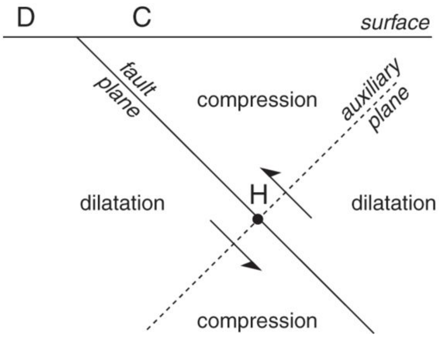


Fig. 7.8 Regions of compression and dilatation around an

earthquake focus, separated by the fault plane and the auxiliary plane.

7.2.5.1 Single-Couple and Double-Couple Radiation Patterns

The amplitudes of P-waves and S-waves vary with distance from their source because of geometric damping and viscoelastic attenuation. The amplitudes also depend on the angle at which the seismic ray leaves the source. This geometric factor can be calculated, assuming a model for the source mechanism. The simplest is to represent the source by a single pair of antiparallel motions. Analysis of the amplitude of the P-wave as a function of the angle θ between a ray and the plane of the fault (Fig. 7.9a) gives an equation of the form

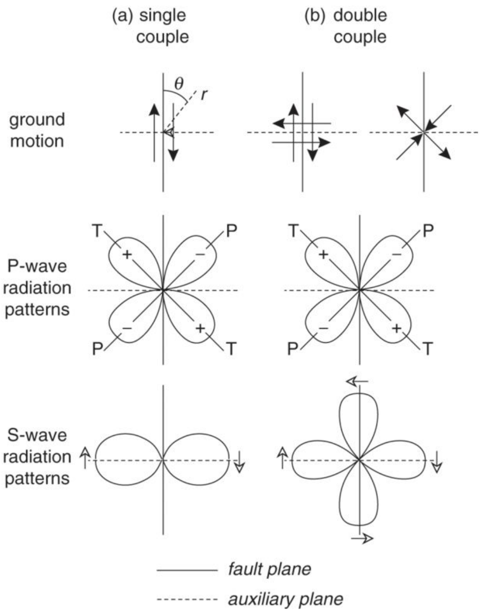

(a) single couple
(b) double couple
ground motion
θ
r
P-wave radiation patterns
T
P
+
−
S-wave radiation patterns
fault plane
auxiliary plane

Fig. 7.9 Azimuthal patterns of amplitude variation for (a) single-couple and (b) double-couple earthquake source models are the same for P-waves but differ for S-waves. The radius of each pattern at azimuth θ from the fault plane is proportional to the amplitude of the seismic wave in this direction. For P-waves the fault plane and auxiliary plane are nodal planes of zero displacement. The maximum compressional amplitude is along the T-axis at an angle

of 45° to the fault plane. The dilatational amplitude is maximum along the P-axis, also at 45° to the fault plane.

$$A(r,t,\alpha,\theta) = A_0(r,t,\alpha)\sin^2 2\theta, \tag{7.2}$$

in which $A_0(r, t, \alpha)$ describes the decrease of amplitude with distance r, time t, and seismic P-wave velocity α. A plot of the amplitude variation with θ is called the radiation pattern of the P-wave amplitude, which for the single-couple model has a quadrupolar character (Fig. 7.9a). It consists of four lobes, two corresponding to the angular variation of amplitude where the first motion is compressional, and two where the first motion is dilatational. The lobes are separated by the fault plane and the auxiliary plane.

The radiation pattern for S-waves from a single-couple source is described by an equation of the form

$$B(r,t,\beta,\theta) = B_0(r,t,\beta)\sin^2\theta, \tag{7.3}$$

where the amplitude B is now dependent on the S-wave velocity β. The radiation pattern has a dipolar character consisting of two lobes in which the first motions are of opposite sense.

An alternative model of the earthquake source is to represent it by a *pair* of orthogonal couples (Fig. 7.9b). The double-couple source gives the same form of radiation pattern for P-waves as the single-couple source, but the radiation pattern for S-waves is quadrupolar instead of dipolar. This difference in the S-wave characteristics enables the seismologist to determine which of the two earthquake source models is applicable. S-wave observations of tectonic earthquakes clearly show that their source is consistent with the double-couple model .

Note that the maximum P-wave amplitudes occur at 45° to the fault plane. The directions of maximum amplitude in the compressional and dilatational fields define the T-axis and P-axis, respectively. Here, T and P imply "tension" and "compression," respectively, the stress conditions before faulting. Geometrically the P- and T-axes are the bisectors of the angles between the fault plane and auxiliary plane. The orientations of these axes and of the fault plane and auxiliary plane can be obtained even for distant earthquakes by analyzing the directions of first motions recorded in seismograms of the events. The analysis is called a fault plane solution, or focal mechanism solution.

7.2.5.2 Fault Plane Solutions

The ray path along which a P-wave travels from an earthquake to the seismogram is curved because of the variation of seismic velocity with depth. The first step in the fault plane solution is to trace the ray back to its source. A fictitious small sphere is imagined to surround the focus (Fig. 7.10a) and the point at which the ray intersects its surface is computed with the aid of P-wave velocity models of the Earth. The azimuth and dip of the angle of departure of the ray from the earthquake focus are calculated and plotted as a point on the lower hemisphere of the small sphere. This direction is then projected onto the horizontal plane through the epicenter. The projection of the entire lower hemisphere is called a stereogram. The direction of the ray is marked with a solid point if the first motion was a push away from the focus (i.e., the station lies in the field of compression). An open point indicates that the first motion was a tug toward the focus (i.e., the station lies in the field of dilatation).

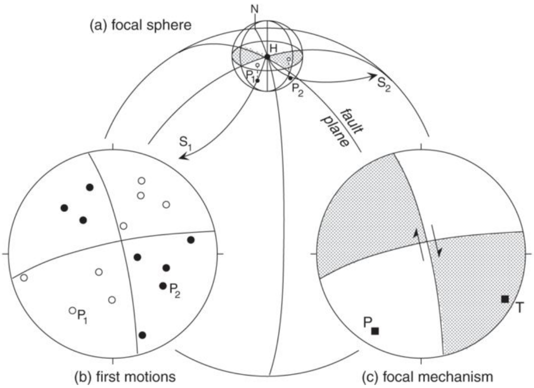


Fig. 7.10 Method of determining the focal mechanism of an earthquake. (a) The focal sphere surrounding the earthquake focus, with two rays S_1 and S_2 that cut the sphere at P_1 and P_2, respectively. (b) The points P_1 and P_2 are plotted on a lower-hemisphere stereogram as first-motion pushes (solid points) or tugs (open points). (c) Beachball plot of the regions of compression (shaded) and tension (unshaded). The P- and T-axes are located on the bisectors of the angles between the fault plane and auxiliary plane.

First-motion data of any event are usually available from several seismic stations that lie in different directions from the focus. The solid and open points on the stereogram usually fall in distinct fields of compression and dilatation (Fig. 7.10b). Two mutually orthogonal planes are now drawn so as to delineate these fields as well as possible. The two mutually orthogonal planes correspond to the fault plane and the auxiliary plane, although it is not possible to decide which is the active fault plane on the seismic data alone. The regions of the stereogram corresponding to compressional first motions are usually shaded to distinguish them from the regions of dilatational first motions (Fig. 7.10c). The result is commonly referred to as a *beachball plot*. The P- and T-axes are the lines that bisect the angles between the fault plane and auxiliary plane in the fields of dilatation and compression, respectively. To attach a physical meaning to the P- and T-axes we will have to take a closer look at the mechanics of faulting.

7.2.5.3 Mechanics of Faulting

To understand the relation between radiation patterns of seismic waves and the actual state of stress in the Earth that leads to material failure, we first consider the direction where stress attains its maximum value σ_{max}. By definition, a stress is positive for tension. Therefore, the extension of the material is maximal in the direction of σ_{max}. Perpendicular to the direction of maximum extension, the material experiences maximal compression. This corresponds to the minimum value of stress, σ_{min}.

The theory of faulting of homogeneous materials has been developed by studying the failure of materials under *compressional stress* (pressure), which is directed inward toward the origin of the coordinate axes. The minimum tensional stress σ_{min} corresponds to the maximum compressional stress, and vice versa. The reason for taking this view is that geologists are interested in the behavior of materials within the Earth, where pressure builds up with increasing depth and faulting occurs under high confining pressures.

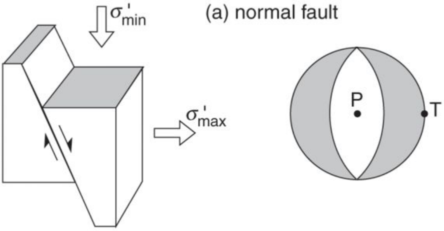
We can combine both points of view if we consider stress to consist of a part that causes change of volume and a part that causes distortion. The first of these is called the *lithostatic stress* (or *hydrostatic stress* in the case of fluids) and is defined as the mean (σ_{m}) of the three diagonal stress components: $\sigma_{\mathrm{m}} = (\sigma_{xx} + \sigma_{yy} + \sigma_{zz})/3$. If we now subtract this value from each of the principal stresses we get the *deviatoric stresses:* σ'_{min} and σ'_{max}. Note that σ'_{max} is positive; it is a tensional stress, directed outward. However, σ'_{min} is negative; it is a compressional stress, directed inward.

Failure of a material occurs on the plane of maximum *shear stress*. For a perfectly homogeneous material this is the plane oriented at 45° to the axes of maximum and minimum compressional stress, i.e., it bisects the angle between the σ'_{min} and σ'_{max} axes. In real materials inhomogeneity and the effects of internal friction result in failure on a fault plane inclined at 20°–30° to the axis of maximum compression. Seismologists often ignore this complication when they make fault plane analyses. The axis of maximum tensional stress $\sigma'_{\max}$ is often equated with the T-axis, in the field of compressional first motions on the bisector of the angle between the two principal planes. The axis of maximum compressional stress $\sigma'_{\min}$ is equated with the P-axis, in the field of dilatational first motions. In reality the direction of $\sigma'_{\min}$ will lie between the P-axis and the fault plane.

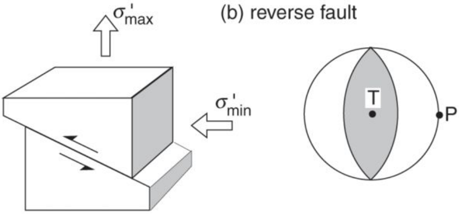
The locations of P and T may at first seem strange, for the axes appear to lie in the wrong quadrants. However, one must keep in mind that the orientations of the principal stress axes correspond to the stress pattern *before* the earthquake, while the fault plane solution shows the ground motions that occurred *after* the earthquake. The focal mechanism analysis makes it possible to interpret the directions of the principal axes of stress in the Earth's crust that led to the earthquake.

There are three main types of tectonic fault. These can be distinguished by the orientations of the principal axes of stress to the horizontal plane (Fig. 7.11). The focal solutions of earthquakes associated with each type of fault have characteristic geometries. When motion on the fault occurs up or down the fault plane, it is called a *dip–slip* fault; when the motion is horizontal, parallel to the strike of the fault, it is called a *strike–slip* fault.

(a) normal fault



(b) reverse fault



(c) strike–slip fault

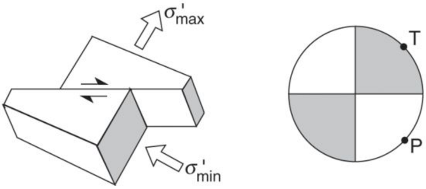


Fig. 7.11 The three main types of fault and their focal mechanisms. *Left*: the orientations of each fault plane and the maximum and minimum deviatoric stresses, σ'_{max} and σ'_{min}. *Right*: focal mechanisms and orientations of P- and T- axes.

Two classes of dip–slip fault are distinguished, depending on the sense of the vertical component of motion. In a *normal fault*, the block on the upper side of the fault drops down an inclined plane of constant steepness relative to the underlying block (Fig. 7.11a). The corresponding fault plane solution has regions of compression at the margins of the stereogram. The T-axis is horizontal and the P-axis is vertical. A special case of this type is the *listric fault*, in which the steepness of the fault surface is not constant but decreases with increasing depth.

In the second type of dip–slip fault, known as a *reverse fault* or *thrust fault*, the block on the upper side of the fault moves up the fault plane, overriding the underlying block (Fig. 7.11b). The fault plane solution is typified by a central compressional sector. The orientations of the axes of maximum tension and compression are the inverse of the case for a normal fault. The T-axis is now vertical and the P-axis is horizontal. When the fault plane is inclined at a very flat angle, the upper block can be transported over large horizontal distances. This special type of *overthrust fault* is common in regions of continental collision, as for example in the Alpine–Himalayan mountain belts.

The simplest type of strike–slip fault is the *transcurrent fault*, in which the fault plane is steep or vertical (Fig. 7.11c). To cause this type of fault the T- and P- axes must both lie in the horizontal plane. The fault plane solution shows two compressional and two dilatational quadrants. Each side of the fault moves in opposite horizontal directions. If the opposite side of the fault to an observer is perceived to move to the left, the fault is said to be *sinistral*, or left-handed; if the opposite side moves to the right, the fault is *dextral*, or right-handed.

A variant of the strike–slip fault plays a very important role in the special context of plate tectonics. A *transform fault* allows horizontal motion of one plate relative to its neighbor. It joins segments of a spreading ocean ridge, or segments of a subducting plate margin. It is important in plate tectonics because it constitutes a conservative plate margin at which the tectonic plates are neither formed nor destroyed. The relative motion on the transform fault therefore reveals the direction of motion on adjacent segments of an active plate margin. The sense of motion is revealed by the pattern of compressional and dilatational sectors on the fault plane solution.

7.2.5.4 Focal Mechanisms at Active Plate Margins

Some of the most impressive examples of focal mechanism solutions have been obtained from active plate margins. The results fully confirm the expectations of plate tectonic theory and give important evidence for the directions of plate motions. We can first ask what types of focal mechanism should be observed at each of the three types of active plate margin. In the theory of plate tectonics these are the *constructive* (or “accreting”), *conservative* (or “transform”), and *destructive* (or “consuming”) margins.

Oceanic spreading systems consist of both constructive and conservative margins. The seismicity at these plate margins forms narrow belts on the surface of the globe. The focal depths are predominantly shallow, mostly less than 10 km below the ocean bottom. Active ridge segments are separated by transform faults (Fig. 7.12). New oceanic lithosphere is generated at the spreading oceanic ridges; the separation of the plates at the spreading center is accompanied by extension. The plates are pulled apart by the plate tectonic forces. The extensional nature of the ridge tectonics is documented by fault plane solutions indicative of normal faulting, as is seen for some selected earthquakes along the Mid-Atlantic Ridge (Fig. 7.13). In each case the fault plane is oriented parallel to the strike of the ridge. On a ridge segment that is nearly normal to the nearest transform faults the focal mechanism solution is symmetric, with shaded compressional quadrants at the margins of the stereogram. Note that where the ridge is inclined to the strike of the transform fault the focal mechanism solution is not symmetric. This means that the plates are not being pulled apart perpendicular to the ridge. The fault plane is still parallel to the strike of the ridge, but the slip vector is oblique; the plate motion has a component perpendicular to the ridge and a component parallel to the ridge. We can understand why the direction of plate motion is not determined by the strike of the ridge axis by examining the motion on the adjacent transform faults.

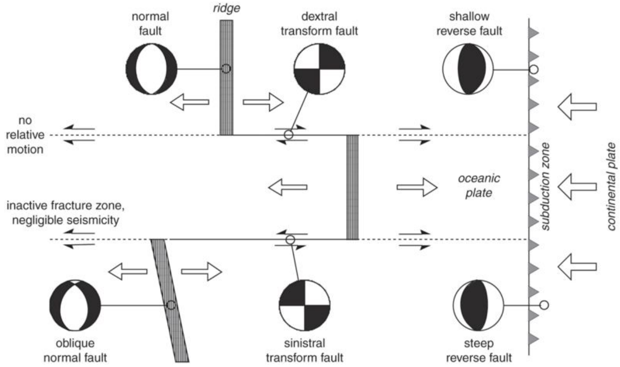

ridge
normal fault
dextral transform fault
shallow reverse fault
no relative motion
oceanic plate
subduction zone
continental plate
inactive fracture zone, negligible seismicity
oblique normal fault
sinistral transform fault
steep reverse fault

Fig. 7.12 Fault plane solutions for hypothetical earthquakes at an ocean ridge and transform fault system. Note that the sense of movement on the fault is not given by the apparent offset of the ridge. The focal mechanisms of earthquakes on the transform fault reflect the relative motion between the plates. Note that in this and similar figures the sector with *compressional* first motions is shaded.

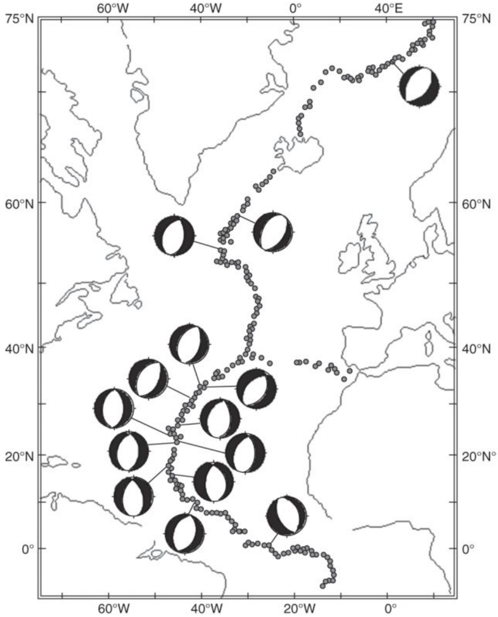


Fig. 7.13 Fault plane solutions for earthquakes along the oceanic Mid-Atlantic Ridge, showing the prevalence of extensional tectonics with normal faulting in the axial zone of the spreading center.

Data source: P.Y. Huang, S.C. Solomon, E.A. Bergman and J.L. Nabelek, Focal depths and mechanisms of Mid-Atlantic Ridge earthquakes from body waveform inversion, *J. Geophys. Res.* 91, 579–598, 1986.

The special class of strike–slip fault that joins active segments of a ridge or subduction zone is called a transform fault because it marks a conservative plate margin, where the lithospheric plates are being neither newly generated nor destroyed. The adjacent plates move past each other on the active fault. Relative plate motion is present only between the ridge segments. Almost the entire seismicity on the transform fault is concentrated in this region. On the parts of the fracture zone outside the segment of active faulting the plates move parallel to each other and there is little or no seismicity.

Because the relative motion is horizontal, the fault plate solution is typical of a strike–slip fault. However, if the visible offset of the ridge segments were used to interpret the sense of motion on this fault, the wrong conclusion would be drawn. The conventional interpretation of this class of faults as a transcurrent fault was an early stumbling block to the development of plate tectonic theory. As indicated by arrows on the focal mechanism diagrams (Fig. 7.12), the relative motion on a transform fault is opposite to what one would expect for a transcurrent fault. It is determined by the opposite motions of the adjacent plates and not by the offset of the ridge segments. Hence, the focal mechanisms for a number of earthquakes on transform faults at the Mid-Atlantic Ridge in the central Atlantic reflect the eastwards motion of the African plate and westwards motion of the American plate (Fig. 7.14).

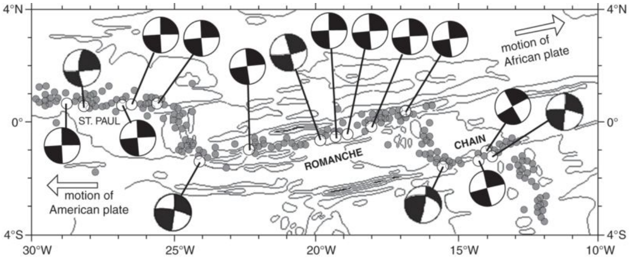


Fig. 7.14 Fault plane solutions for earthquakes on the St. Paul, Romanche, and Chain transform faults in the Central Atlantic ocean. Most focal mechanisms show right lateral (dextral) motions on these faults, corresponding to the relative motion between the African and American plates.

Source: J.F. Engeln, D.A. Wiens and S. Stein, Mechanisms and depths of Atlantic transform earthquakes, *J. Geophys. Res.* 91, 548–577, 1986

If we ignore changes in plate motion (which can, however, occur sporadically on some ridge systems), the offset of the neighboring ridge segments is a permanent feature that reflects how the plates first split apart. The orientations of the transform faults are very important, because the plates must move parallel to these faults. Thus the transform faults provide the key to determining the directions of plate motion. Where a ridge axis is not perpendicular to the transform fault the plate motion will have a component parallel to the ridge segment, which gives an oblique focal mechanism on the ridge.

A destructive (or consuming) plate margin is marked by a subduction zone, where a plate of oceanic lithosphere is destroyed by plunging under another plate of oceanic or continental lithosphere. Because this is a margin of convergence of the adjacent plates, the earthquake fault plane solutions are typical of a compressional regime (Fig. 7.11). The regions of compressional first motion are in the center of the stereogram, indicating reverse faulting; the P-axes of maximum compressive stress are perpendicular to the strike of the subduction zone.

The type of focal mechanism observed at a subduction zone is dependent on the focal depth of the earthquake. This is because the state of stress varies within the subducting plate. At first the overriding plate is thrust at a shallow angle over the subducting plate. In the seismic zone at the contact between the two plates, earthquake focal mechanisms are typical of low-angle reverse faulting. The focal mechanisms of earthquakes along the west coast of Mexico illustrate the first of these characteristics (Fig. 7.15). The selected earthquakes have large magnitudes ($6.9 \leq M_s \leq 7.8$) and shallow focal depths. The strikes of the fault planes follow the trend of the oceanic trench along the Mexican coastline. The focal mechanisms have a central sector of compressional first motions and the fault plane is at a low angle to the northeast, typical of overthrusting tectonics. The seismicity pattern documents the subduction of the Cocos plate under the North American plate.

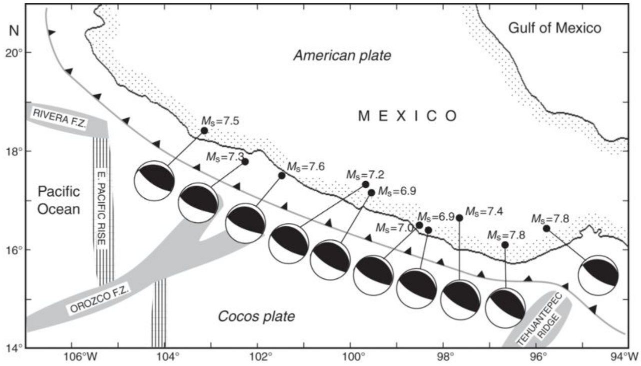

N
20°
18°
16°
14°
Gulf of Mexico
American plate
M E X I C O
RIVERA F.Z.
E. PACIFIC RISE
Pacific Ocean
OROZCO F.Z.
Cocos plate
TEHUANTEPEC RIDGE
M_s=7.5
M_s=7.3
M_s=7.6
M_s=7.2
M_s=6.9
M_s=6.9
M_s=7.4
M_s=7.0
M_s=7.8
M_s=7.8
106°W
104°
102°
100°
98°
96°
94°W

Fig. 7.15 Fault plane solutions for selected large shallow earthquakes in the subduction zone along the west coast of Mexico. The focal mechanisms indicate low-angle overthrusting, as the Cocos plate is subducted to the northeast under Mexico.

Source: S.K. Singh, T. Dominguez, R. Castro and M. Rodriguez, P-waveform of large shallow earthquakes along the Mexico subduction zone, *Bull. Seism. Soc. Amer.* 74, 2135–2156, 1984.

As a result of the interplate collision, the subducting oceanic plate is bent downwards and its state of stress changes. Deeper than 60–70 km the seismicity does not arise from the contact between the converging plates. It is caused by the stress pattern within the subducted plate itself. The focal mechanisms of some intermediate-depth earthquakes (70–300 km) show down-dip extension (i.e., the T-axes are parallel to the surface of the dipping slab), but some show down-dip compression (i.e., the P-axes are parallel to the dip of the slab). At great depths in most Wadati–Benioff zones the focal mechanisms indicate down-dip compression.

7.2.5.5 Focal Mechanisms in Continental Collisional Zones

When the continental portions of converging plates collide, they resist subduction into the denser mantle. The deformational forces are predominantly horizontal and lead to the formation of folded mountain ranges. The associated seismicity tends to be diffuse, spread out over a large geographic area. The focal mechanisms of earthquakes in the folded mountain belt reflect the ongoing collision. The collision of the northward-moving Indian plate with the Eurasian plate in the Late Tertiary led to the formation of the Himalayas. Focal mechanisms of earthquakes along the arcuate mountain belt show that the present style of deformation consists of two types (Fig. 7.16). Two fault plane solutions south of the main mountain belt correspond to an extensional regime with normal faulting. To the north, in southern Tibet, the fault plane solutions also show normal faulting on north–south-oriented fault planes. Beneath the Lesser Himalaya, seismic events distributed along the entire 1800 km length of the mountain chain indicate a deformational regime with some strike–slip faulting but predominantly low-angle thrust faulting. The Indian continental crust appears to be thrusting at a shallow angle under the Tibetan continental crust. This causes crustal thickening toward the north. In the main chain of the Lesser Himalaya the *minimum* compressive stress is vertical. The increased vertical load due to crustal thickening causes the directions of the principal stresses to change, so that under southern Tibet the *maximum* compressive stress is vertical.

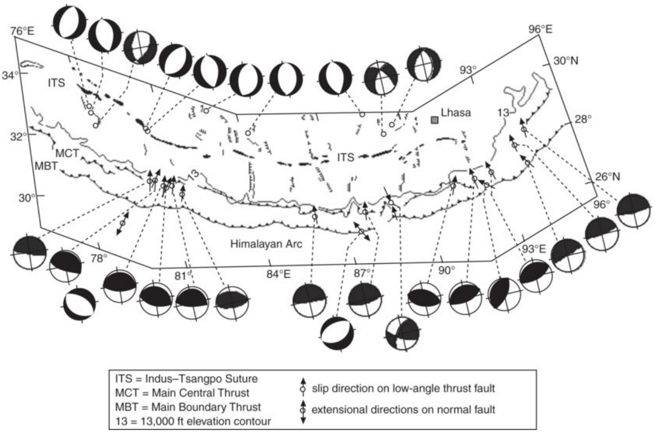

76°E
96°E
34°
ITS
30°N
93°
Lhasa
13
28°
32°
MCT
MBT
ITS
13
26°N
30°
96°
Himalayan Arc
93°E
78°
81°
84°E
87°
90°
ITS = Indus–Tsangpo Suture
MCT = Main Central Thrust
MBT = Main Boundary Thrust
13 = 13,000 ft elevation contour
slip direction on low-angle thrust fault
extensional directions on normal fault

Fig. 7.16 Fault plane solutions of earthquakes along the arcuate Himalayan mountain belt show normal faulting on north–south-oriented fault planes in southern Tibet, and mainly low-angle thrust faults along the Lesser Himalaya mountain chain.

Source: J. Ni and M. Barazangi, Seismotectonics of the Himalayan collision zone: Geometry of the underthrusting Indian plate beneath the Himalaya, *J. Geophys. Res.* 89, 1147–1163, 1984.

A different type of collisional tectonics is shown by fault plane solutions from the Alpine mountain belt in south-central Europe. The Alps were formed during the collision between the African and European plates, starting in the Early Tertiary. The focal mechanisms of many, mostly small earthquakes show that they are predominantly associated with strike–slip faults (Fig. 7.17). The horizontal projections of the compressional axes are oriented almost perpendicular to the strike of the Alps and rotate along the arc. The fault plane solutions for the modern seismicity indicate that the Alpine fold-belt is a region of continuing interplate collision.

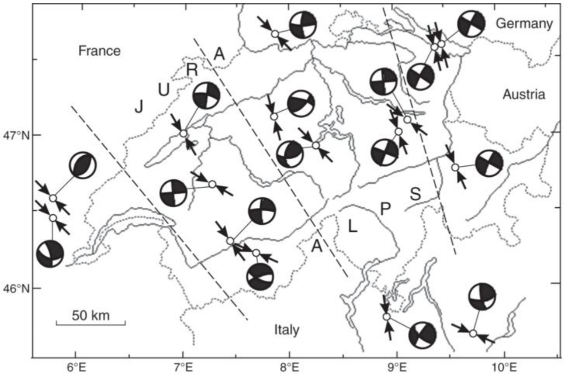


Fig. 7.17 Fault plane solutions for earthquakes in or near the Swiss Alps in central Europe. The arrows show the horizontal components of the interpreted axes of maximum compression. They are approximately at right angles to the strike of the mountain ranges of the Jura and Alps.

Sources: N. Pavoni, Erdbeben im Gebiet der Schweiz, *Eclogae Geol. Helv.* 70/2, 351–371, 1977; D. Mayer Rosa and S. Müller, Studies of seismicity and selected focal mechanisms of Switzerland, *Schweiz. miner. petr. Mitt.* 59, 127–132, 1979.

7.2.6 Earthquake Size

There are two methods of describing how large an earthquake is. The *intensity* of the earthquake is a subjective parameter that is based on an assessment of visible effects. It therefore depends on factors other than the actual size of the earthquake. The *magnitude* of an earthquake is determined instrumentally and is a more objective measure of its size, but it says little directly about the seriousness of the ensuing effects. It is usually the magnitude that is reported in news coverage of a major earthquake, whereas the intensity is a more appropriate parameter for describing the severity of its effects on mankind and the environment.

7.2.6.1 Earthquake Intensity

Large earthquakes produce alterations to the Earth's natural surface features, or severe damage to man-made structures such as buildings, bridges, and dams. Even small earthquakes can result in disproportionate damage to these edifices when inferior construction methods or materials have been utilized. The intensity of an earthquake at a particular place is classified on the basis of the local character of the visible effects it produces. It depends very much on the acuity of the observer, and is in principle subjective. Yet, intensity estimates have proved to be a viable method of assessing earthquake size, including historical earthquakes.

The first attempt to grade earthquake severity was made in the late eighteenth century by Domenico Pignataro, an Italian physician who classified more than 1000 earthquakes that affected the southern Italian province of Calabria in the years 1783–1786. His crude analysis classified the earthquakes according to whether they were very strong, strong, moderate, or slight. In the mid-nineteenth century an Irish engineer, Robert Mallet, produced a list of 6831 earthquakes and plotted their estimated locations, producing the first map of the world's seismicity and establishing that earthquakes occurred in distinct zones. He also used a four-stage intensity scale to grade earthquake damage, and constructed the first isoseismal maps with lines that showed areas with broadly equal grades of damage. The Rossi–Forel intensity scale, developed in the late nineteenth century by the Italian scientist Michele Stefano Conte de Rossi and the Swiss scientist François Alphonse Forel, incorporated ten stages describing effects of increasing damage. In 1902 an Italian volcanologist, Giuseppe Mercalli, proposed a still more extensive, expanded intensity scale which reclassified earthquake severity in 12 stages. A variation, the modified Mercalli (MM) scale, was developed in 1931 to suit building conditions in the USA, where a later modification is in common use. The Medvedev–Sponheuer–Karnik (MSK) scale, introduced in Europe in 1964 and modified in 1981, also has 12 stages and differs from the MM scale mainly in details. A new European Macroseismic Scale was adopted in 1998 (EMS-98); an abridged version is shown in Table 7.1. The new 12-stage EMS scale is based on the MSK scale but takes into account the vulnerability of buildings to earthquake damage and incorporates more rigorous evaluation of the extent of damage to structures with different building standards.

Table 7.1 Abridged and simplified version of the European Macroseismic Scale 1998 (EMS-98) for earthquake intensity

The scale focuses especially on the effects on people and buildings. It takes into account classifications of both the vulnerability of a structure (i.e., the materials and method of construction) and the degree of damage.

Intensity	Description of effects
I–IV	*Light to moderate earthquakes*
I	**Not felt.**
II	**Scarcely felt**: Felt only by a few individuals at rest in houses.
III	**Weak**: Felt indoors by a few people. People at rest feel a swaying or light trembling.

IV	**Largely observed**: Felt indoors by many people; outdoors by very few. A few people are awakened. Windows, doors, and dishes rattle.
V–VIII	*Moderate to severe earthquakes*
V	**Strong**: Felt indoors by most, outdoors by few. Many sleeping people awake. A few are frightened. Buildings tremble throughout. Hanging objects swing considerably. Small objects are shifted. Doors and windows swing open or shut.
VI	**Slightly damaging**: Many people are frightened and run outdoors. Some objects fall. Many houses suffer slight non-structural damage like hair-line cracks and fall of small pieces of plaster.
VII	**Damaging**: Most people are frightened and run outdoors. Furniture is shifted and objects fall from shelves in large numbers. Many well-built ordinary buildings suffer moderate damage: small cracks in walls, fall of plaster, parts of chimneys fall down; older buildings may show large cracks in walls and failure of fill-in walls.
VIII	**Heavily damaging**: Many people find it difficult to stand. Many houses have large cracks in walls. A few well-built ordinary buildings show serious failure of walls, while weak older structures may collapse.
IX–XII	*Severe to destructive earthquakes*
IX	**Destructive**: General panic. Many weak constructions collapse. Even well-built ordinary buildings show very heavy damage: serious failure

	of walls and partial structural failure.
X	**Very destructive**: Many ordinary well-built buildings collapse.
XI	**Devastating**: Most ordinary well-built buildings collapse, even some with good earthquake-resistant design are destroyed.
XII	**Completely devastating**: Almost all buildings are destroyed.

Source: European Seismological Commission, 1998.

In order to evaluate the active seismicity of a region, questionnaires may be distributed to the population, asking for observations that can be used to estimate the intensity experienced. The questionnaires are evaluated with the aid of an intensity scale, and the intensity recorded at the location of each observer is plotted on a map. Continuous lines are then drawn to outline places with the same intensity, in the same way that contour lines are used on topographic maps to show elevation. For earthquakes that occurred in the pre-instrumental era of seismology, other sources of information – such as newspaper reports – may be used to estimate intensity maps (Fig. 7.18). This makes the concept of earthquake intensity particularly valuable for the study of historic seismicity.

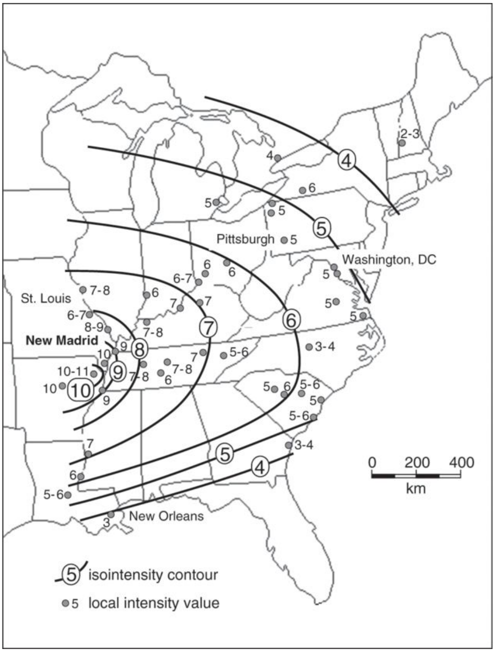

2-3
4
4
6
5
5
5
Pittsburgh 5
Washington, DC
5
6
6-7
6
6
St. Louis
7-8
6
7
7
5
6-7
8-9
7-8
7
6
5
New Madrid
9
8
10
3-4
5-6
7
10-11
9
7-8
7-8
6
10
10
9
5
6
5-6
5
5-6
3-4
7
5
4
0 200 400
km
6
5-6
3
New Orleans
5 isointensity contour
5 local intensity value

Fig. 7.18 Isoseismal map with contours of equal intensity for the New Madrid, Missouri, earthquake of 1811.

Source: O.W. Nuttli, The Mississippi Valley earthquakes of 1811 and 1812: Intensities, ground motion and magnitudes, *Bull. Seism. Soc. Amer.* 63, 227–248, 1973.

Comparison of the *isoseismal* maps with geological maps helps explain the response of the ground to an earthquake. This is valuable information for understanding earthquake risk. The foundation on which structures are erected plays a vital role in their survival of an earthquake. For example, soft sediments can amplify ground motion, enhancing the damage caused. This is even more serious when the sediments have a high water content, in which case liquefaction of the sediments can occur, robbing structures built on them of support and promoting their collapse.

There are numerous examples of this having occurred. In the great Alaskan earthquake in 1964 a section of Anchorage that was built on a headland underlain by a wet clay substratum collapsed and slid downslope to the sea. In 1985 a large earthquake with magnitude

8.1 struck the Pacific coast of Mexico. About 350 km away in Mexico City, despite the large distance from the epicenter, the damage to buildings erected on the alluvium of a drained lakebed was very severe, while buildings set on a hard rock foundation on the surrounding hills suffered only minor damage. In the Loma Prieta earthquake of 1989 severe damage was caused to houses built on landfill in San Francisco's Mission district, and an overhead freeway built on pillars on young alluvium north of Oakland collapsed dramatically. Both regions of destruction were more than 70 km from the epicenter in the Santa Cruz mountains. Similarly, in the San Francisco earthquake of 1906 worse damage occurred to structures built on landfill areas around the shore of the bay than to those with hard rock foundations in the hills of the San Francisco peninsula. More recently, the 28 September 2018 Sulawesi earthquake caused massive soil liquefaction, resulting in the destruction of hundreds of houses in and around the city of Palu.

7.2.6.2 Earthquake Magnitude

Magnitude is an experimentally determined measure of the size of an earthquake. In 1935 Charles Francis Richter attempted to grade the sizes of local earthquakes in Southern California on the basis of the amplitude of the ground vibrations they produced at a known distance from the epicenter. Richter's original definition of magnitude was based on surface wave amplitudes (A_S) recorded at an epicentral distance of 100 km. Because seismometers were located at various distances from the earthquake, an extra term was added to compensate for attenuation of the wave with increasing epicentral distance. Increasingly sensitive instruments allowed the recording of signals from distant earthquakes; those from events with epicentral distances greater than 20° are known as *teleseismic* signals. Originally, the magnitude was determined from the horizontal ground motion, because seismological stations were equipped mainly with horizontal-motion seismometers. However, the surface waves recorded by these instruments consist of superposed Love and Rayleigh waves, which complicates the interpretation of the records. Vertical-motion seismometers record only the Rayleigh waves (together with P- and SV-waves), and so progressively the definition of surface wave magnitude has come to be based on the vertical component of motion. The majority of surface wave magnitudes assigned to earthquakes worldwide are now based on vertical-motion records.

The International Association for Seismology and Physics of the Earth's Interior (IASPEI) has adopted the following definition of the surface wave magnitude (M_s) of an earthquake:

$$M_s = \log_{10}\left(\frac{A_s}{T}\right)_{\max} + 1.66\log_{10}(\Delta) + 3.3, \quad (7.4)$$

where A_s is the vertical component of the ground motion in microns (μm) determined from the maximum Rayleigh wave amplitude, T is the period of the wave (required to be 18–22 s), Δ is the epicentral distance in degrees ($20^\circ \le \Delta \le 160^\circ$), and where the earthquake has a focal depth of less than 50 km. An equation similar to Eq. (7.4) applies to broadband recordings; in this case $(A_s/T)_{\max}$ corresponds to the maximum ground velocity.

The depth of the source affects the nature of the seismic wave train, even when the same energy is released. An earthquake with a deep focus may generate only a small surface wave train, while shallow earthquakes cause very strong surface waves. Equation (7.4) for M_s was derived from the study of shallow earthquakes, observed at a distance greater than 20°. Therefore, corrections must be made to the computed value of M_s to compensate for the effects of a focal depth greater than 50 km or epicentral distance less than 20°.

The amplitude of body waves is much less sensitive to the focal depth. As a result, earthquake magnitude scales have also been developed for use with body waves. An equation, proposed by Beno Gutenberg in 1945, can be used to calculate a *body wave magnitude*

(m_b) from the maximum amplitude (A_p) of the ground motion associated with P-waves having a period (T) of less than 3 s:

$$m_b = \log_{10}\left(\frac{A_p}{T}\right) + Q(\Delta, h), \tag{7.5}$$

where $Q(\Delta, h)$ is an empirical correction for signal attenuation due to epicentral distance (Δ) and focal depth (h) that is made by reading it directly from a graph or looking it up in a table.

For some earthquakes, both M_s and m_b can be calculated. Unfortunately, the different magnitude estimates often do not agree well, except for small earthquakes. This is due to the way the ground responds to a seismic event, and to the different natures of body waves and surface waves. Body waves have a different dependence of amplitude on frequency than do surface waves. While m_b is estimated from a high-frequency (around 1 Hz) phase, M_s is determined from low-frequency (around 0.05 Hz) vibrations. Above a certain size, each method becomes insensitive to the size of the earthquake, and exhibits *magnitude saturation*. This occurs for body wave estimates at around $m_b = 6$; all larger earthquakes give a similar body wave magnitude. Similarly, surface wave estimates of magnitude saturate at $M_s = 8$. Thus, for very large earthquakes, M_s and m_b underestimate the energy released. An alternative definition of magnitude, based upon the long-period spectrum of the seismic wave, is preferred for very large earthquakes. It makes use of the physical dimensions of the focus.

As discussed in the elastic rebound model (Section 7.2.2), a tectonic earthquake arises from abrupt displacement of a segment of a fault. The area S of the fractured segment and the amount D by which it slipped can be inferred. Together with the rigidity modulus μ of the rocks adjacent to the fault, these quantities define the *seismic moment* M_0 of the earthquake. Assuming that the displacement and rigidity are constant over the area of the rupture,

$$M_0 = \mu S D. \tag{7.6}$$

The seismic moment can be used to define a *moment magnitude* (M_W). The definition adopted by the responsible commission of IASPEI is:

$$M_w = \frac{2}{3}(\log_{10} M_0 - 9.1). \tag{7.7}$$

In this equation M_0 is in N m. If c.g.s. units are used instead of SI units, M_0 is expressed in dyne cm and the corresponding equation is:

$$M_w = \frac{2}{3}(\log_{10} M_0 - 16.1). \tag{7.8}$$

M_W is more appropriate for describing the magnitudes of very large earthquakes. It has largely replaced M_s in the scientific evaluation of earthquake size, although M_s is often quoted in reports in the media. The moment-magnitudes and surface wave magnitudes of some historical earthquakes are listed in Table 7.2.

Table 7.2 Some important historical earthquakes, with their surface wave magnitudes M_s, moment magnitudes M_w, and the estimated numbers of fatalities

Year	**Epicenter**	**Magnitude**		**Fatalities**
		M_s	**M_W**	
1906	San Francisco, California	8.3	7.8	3000
1908	Messina, Italy	7.2	–	70,000
1923	Kanto, Japan	8.2	7.9	143,000
1952	Kamchatka, Russia	8.2	9.0	
1957	Andreanof Islands,	8.1	8.6	

	Alaska			
1960	Valdivia, Chile	8.5	9.5	5700
1960	Agadir, Morocco	5.9	5.7	10,000
1964	Prince William Sound, Alaska	8.6	9.2	125
1970	Chimbote, Peru	7.8	7.9	66,000
1975	Haicheng, China	7.4	7.0	300
1976	Tangshan, China	7.8	7.5	243,000
1980	El Asnam, Algeria	7.3	–	2590
1985	Michoacan, Mexico	8.1	8.0	9500
1989	Loma Prieta, California	7.1	6.9	63
1994	Northridge, California	6.8	6.7	60
1999	Ismit, Turkey	–	7.6	17,100
2004	Sumatra–Andaman Islands	–	9.1	250,000
2010	Haiti	–	7.0	222,570
2011	Tohoku, Japan	–	9.1	20,000
2016	Muisne, Ecuador	–	7.8	700

The magnitude scale is, in principle, open-ended. Negative magnitudes are possible for small earthquakes because of the logarithms in the magnitude definitions, but very small earthquakes may be difficult to detect because of the presence of noise. The maximum possible magnitude is limited by the shear strength of the crust and upper mantle, and since the beginning of instrumental recording none has been observed with a surface wave magnitude M_s

as high as 9. However, this is also due to the saturation effect explained before. Seismic moment magnitudes M_W of 9 and larger have been computed for some giant earthquakes (Table 7.2). The largest in recorded history was the 1960 Chile earthquake with $M_W = 9.5$.

When an earthquake occurs on a fault, the ruptured area spreads in size from the point of initial failure, akin to opening a zip fastener. If the fault ruptures along a length L, and the fractured segment has a down-dip dimension w (referred to as the width of the faulted zone), the area S is equal to wL. Assuming that the aspect ratio of faults is constant (i.e., w is proportional to L), then the ruptured area S is proportional to L^2. This is a generalization, because different faults have different aspect ratios. Similarly, if the stress drop in an earthquake is constant, the displacement on the fault, D, can be assumed proportional to L. Together, these assumptions imply that the seismic moment M_0 scales as L^3. Assuming that $S \propto L^2$, the seismic moment scales as $S^{3/2}$. This inference is supported by the correlation between seismic moment and the ruptured area S for many shallow earthquakes (Fig. 7.19).

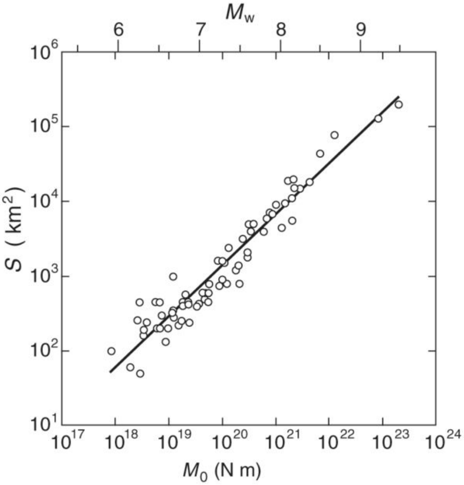


Fig. 7.19 Correlation of the ruptured area (S, in km^2) with the seismic moment (M_0, in N m) and moment magnitude (M_W) for some shallow .

Source: H. Kanamori and E.E. Brodsky, The physics of earthquakes, *Reports on Progress in Physics*, 67, 1429–1496, 2004.

The length L of the rupture zone determines how long the ground motion lasts in an earthquake. This factor has a significant bearing on the extent of damage to structures during moderate to strong earthquakes. For most earthquakes, the speed of propagation of the rupture is around 75–95% of the shear wave velocity. Assuming that the rupture propagates along the fault at about 2.5–3.0 km s^{-1}, the duration of the ground motion can be estimated for an earthquake of a given magnitude or seismic moment (Fig. 7.20). For example, close to the epicenter of a magnitude 5 earthquake the vibration may last a few seconds, whereas in a magnitude 8 earthquake it might last about 50 s. However, there is a lot of scatter about the ideal correlation. In the great Sumatra earthquake, with magnitude 9.0, the rupture continued for some 500 s.

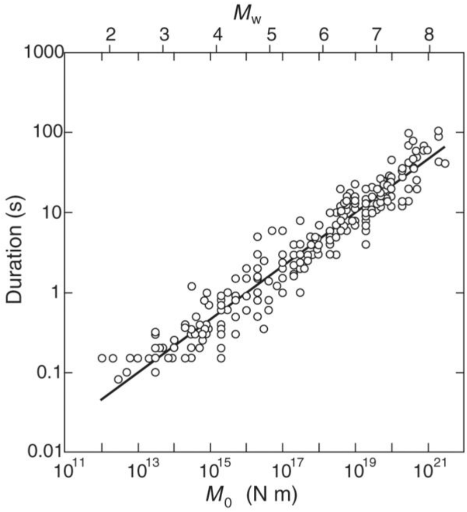


Fig. 7.20 Correlation of the source durations (in seconds) of shallow earthquakes and their seismic moments, M_0 (in N m), and moment magnitudes, M_w.

Source: H. Kanamori and E.E. Brodsky, The physics of earthquakes, *Reports on Progress in Physics*, 67, 1429–1496, 2004.

The seismic moment M_0 of an earthquake is determined in practice from the analysis of seismograms at a large number of seismic stations. The waveform of each seismogram is affected by the epicentral distance, focal depth, and radiation pattern of the earthquake (Section 7.2.5). It is also influenced by source parameters such as the area, orientation, and amount of slip of the ruptured segment, factors which determine the seismic moment M_0. Inversion of the seismogram data leads to an understanding of the earthquake source. At large distances and for long-period components of the seismogram (e.g., $T > 40$ s) that correspond to wavelengths much longer than the dimensions of the faulted area, the source can be considered as a point source. Assuming a radiation model for the source and a velocity model for the propagation, a synthetic seismogram (see also Chapter 6) can be computed for each station and compared with the observed waveform. Iterative adjustment of the source parameters to give a best fit to the waveforms at a number of receivers defines mathematically a *moment tensor*. The tensor can be visualized geometrically as an ellipsoid in which the lengths and orientations of the axes correspond to the moment of the earthquake and the directions of the tensional (T) axis and compressional (P) axis at the source. The location of the system of couples modeled by the moment tensor is the optimum point source location for the earthquake and is called its *centroid*. The *centroid moment tensor* (CMT) analyses of earthquakes are made rapidly and give the source parameters of an earthquake within a few hours. The centroid location may differ from the hypocenter of the earthquake (the place where rupture first occurred). The former is based on analysis of the full seismogram waveform, the latter only on the first arrival times. For example, the centroid of the 2004 Sumatra–Andaman earthquake was located about 160 km west of its epicenter.

7.2.6.3 Relationship Between Magnitude and Intensity

The intensity and magnitude scales for estimating the size of an earthquake are defined independently, but they have some common features. Intensity is a measure of earthquake size based on the extent of local damage it causes at the location of an observer. The definition of magnitude is based on the amplitude of ground motion inferred from the signal recorded by the observer's seismograph, and of course it is the nature of the ground motion – its amplitude, velocity, and acceleration – which produce the local damage used to classify intensity. However, in the definition of magnitude the ground-motion amplitude is corrected for epicentral distance and converted to a focal characteristic. Isoseismal maps showing the regional distribution of damage give the maximum intensity ($I_{\max}$) experienced in an earthquake, which, although influenced by the geographic patterns of population and settlement, is usually near to the epicenter. A moderately strong, shallow-focus earthquake under a heavily populated area can result in higher intensities than a large deep-focus earthquake under a wilderness area (compare, for example, the death tolls for the 1960 Agadir magnitude 5.7 and 1964 Alaskan magnitude 9.2 earthquakes in Table 7.2). However, for earthquakes with focal depth $h < 50$ km the dependence of $I_{\max}$ on the focal depth can be taken into account, and it is possible to empirically find an approximate relation between the maximum intensity and the magnitude (Karnik, 1969):

$$I_{\max} \approx 1.5M_s - 1.8\log_{10}h + 1.7. \tag{7.9}$$

This type of equation is useful for estimating quickly the probable damage that an earthquake causes. For example, it predicts that in the epicentral region of an earthquake with magnitude 5 and a shallow focal depth of 10 km, the maximum MSK intensity will be VII (moderately serious damage), whereas if the focal depth is 100 km, a maximum intensity of only IV–V (minor damage) can be expected.

7.2.7 Energy Released in an Earthquake

The definition of earthquake magnitude is related to the amplitude of a seismic wave. Noting that the energy of a wave is proportional to the square of its amplitude, it should be no surprise that the magnitude is also related to the logarithm of the energy. Several equations have been proposed for this relationship. An empirical formula worked out by Beno Gutenberg (1956) relates the energy release E to the surface wave magnitude M_S:

$$\log_{10}E = 4.8 + 1.5M_s,$$

(7.10)

where E is in joules. The logarithmic nature of the formula means that the energy release increases very rapidly with increasing magnitude. For example, when the magnitudes of two earthquakes differ by 1, their corresponding energies differ by a factor of around 32 (= 101.5). Hence, a magnitude 7 earthquake releases about 1000 (= $10^{(1.5\times2)}$) times the energy of a magnitude 5 earthquake. Another way of regarding this observation is that it takes 1000 magnitude 5 earthquakes to release the same amount of energy as a single large earthquake with magnitude 7. Multiplying the mean number of earthquakes per year by their estimated energy (using one of the energy–magnitude equations) gives an impression of the importance of very large earthquakes. Table 7.3 shows that the earthquakes with $M_S \geq 7$ are responsible for most of the annual seismic energy. In a year in which a very large earthquake ($M_S \geq 8$) occurs, most of the annual seismic energy is released in that single event.

Table 7.3 Average number of earthquakes per year worldwide between 1990 and 2017, except for M ≥ 8, which are averaged since 1900. The mean annual release of seismic energy is estimated with the energy–magnitude relation in Eq. (7.10)

Earthquake	Number per year	Annual energy

magnitude		**(10^{15} J yr^{-1})**
≥ 8.0	≈ 1	≈ 100
7–7.9	14	190
6–6.9	138	45
5–5.9	1508	14
4–4.9	≈ 13,000	4
3–3.9	≈ 130,000	1
2–2.9	≈ 1,300,000	0.4

Data source: US Geological Survey National Earthquake Information Center.

It is rather difficult to appreciate the amount of energy released in an earthquake from the numerical magnitude alone. A few examples help illustrate the amounts of energy involved. Earthquakes with $M_s = 1$ are so weak that they can only be recorded instrumentally; they are referred to as microearthquakes. The energy associated with one of these events is equivalent to the kinetic energy of a medium-sized automobile weighing 1.5 tons traveling at 130 km h^{-1} (80 mph). The energy released by explosives provides another means of comparison, although the conversion of energy into heat, light, and shock waves is proportionately different in the two phenomena. One ton of the explosive tri-nitro-toluene (TNT) releases about 4.2×10^9 joules of energy. Equation (7.10) shows that the 11 kiloton atomic bomb that destroyed Hiroshima released about the same amount of energy as an earthquake with magnitude 5.9. The largest thermonuclear bomb, tested in 1961 by the Soviet Union, released around 210×10^{15} J, approximately equivalent to a magnitude 8.3 earthquake. The 2004 Sumatra earthquake (magnitude 9.0) released an amount of energy equivalent to the detonation of 475 megaton bombs. The world energy consumption in 2017 was around 600×10^{12} J. A similar amount of energy is released by a magnitude 6.7 earthquake.

7.2.8 Earthquake Frequency

Every year there are many small earthquakes, and only a few large ones. According to a compilation published by Beno Gutenberg and Charles Richter in 1954, the mean annual number of earthquakes in the years 1918–1945 with magnitudes 4–4.9 was around 6000, while there were only an average of about 100 earthquakes per year with magnitudes 6–6.9. The empirical relationship between the number of earthquakes above a certain magnitude (N) and magnitude (M_s) is logarithmic and is approximately given by

$$\log_{10} N \approx a - bM_s. \tag{7.11}$$

The value of a varies between about 8 and 9 from one region to another, while b is approximately unity for regional and global seismicity. The mean annual numbers of earthquakes in different magnitude ranges are listed in Table 7.3; the frequency decreases with increasing magnitude (Fig. 7.21a), in accordance with Eq. (7.11). The annual number of large earthquakes with magnitude $M_s \geq 7$ in the years 1900–2017 has varied between extremes of about 10 and 40, but the long-term average is about 20 per year (Fig. 7.21b).

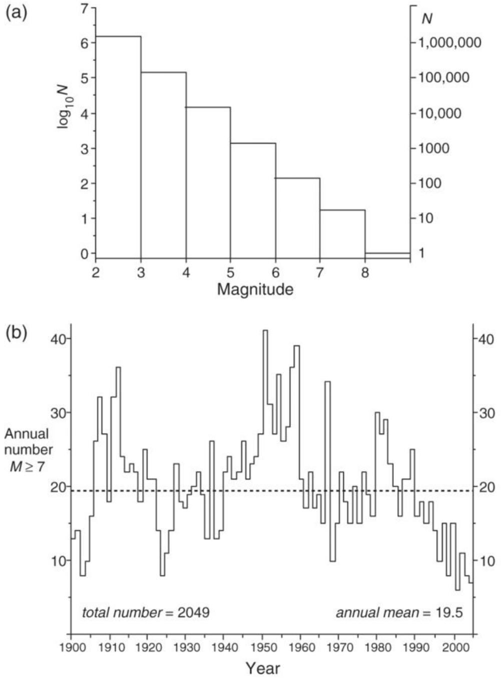


Fig. 7.21 Histograms of (a) the logarithm of the number (*N*) of earthquakes per year with magnitude M_s, and (b) the annual number of earthquakes with magnitudes $M_s \geq 7$ since 1900 (based on data from the US Geological Survey National Earthquake Information Center).

7.3 Earthquakes and Human Society

The relevance of earthquake studies largely derives from their relation to human society. The sudden occurrence of earthquakes motivated countless attempts to predict these potentially destructive events. So far, however, no practically useful method for earthquake prediction has been developed, and the focus has shifted to the long-term analysis of seismic hazard, intended to inform building codes and to prepare the population. In addition to the ground shaking itself, earthquakes can have numerous side-effects, including fires, landslides, and tsunamis. Thanks to the relatively low propagation speed of tsunamis, effective early warnings may be issued.

While global seismicity is dominated by naturally occurring tectonic events, some earthquakes are induced by industrial activity such as oil and gas extraction, the pumping of wastewater into underground formations, and the injection of pressurized fluids into reservoirs in order to enhance the production of hydrocarbons and geothermal energy. Another kind of man-made earthquake is related to nuclear explosions, which may be detected and located using seismic data recorded at long distances.

7.3.1 Earthquake Prediction, Forecasting, and Early Warning

The prediction of destructive earthquakes is among the most challenging problems in geophysics. *Earthquake prediction* is defined as the specification of the timing, location, and magnitude of future earthquakes within stated uncertainties. For a prediction to be significant, these uncertainties must be smaller than those of pure random chance. In the 1970s, it was thought that a practically useful method for earthquake prediction would soon be available. After numerous unsuccessful, or at least questionable, predictions, it has become the consensus that earthquakes cannot be predicted with currently available methods.

Earthquake prediction is distinctly different from *earthquake forecasting*, which is concerned with the statistical assessment of seismic hazard, including the frequency and magnitude of earthquakes in some region over many years or decades. Earthquake forecasting is the basis of codes for the construction of buildings, insurance rates, and awareness and preparedness programs.

In contrast to prediction and forecasting, *earthquake early warning* happens within seconds or minutes after an earthquake has already occurred. Based on the rapid estimation of location and magnitude using seismic waves, an early warning may shut down nuclear power plants, stop high-speed trains, or alert the population. This works because modern communication systems allow an alert to travel faster than seismic waves.

7.3.1.1 Attempts to Predict Earthquakes

The 1975 Haicheng, China, event is the most frequently cited example of an earthquake prediction that was apparently successful. In February 1975, the frequency of minor shocks increased, at first gradually but eventually dramatically, and then there was an ominous pause in earthquake activity. This time gap was interpreted as an indication of an impending earthquake. The population was ordered to vacate their homes several hours before a magnitude 7.4 earthquake that destroyed cities and communities. Although the epicenter was near to the city of Haicheng, the death toll among the 3,000,000 inhabitants of the province was not more than about 2000. Later investigations of the Haicheng prediction criticized its scientific documentation and suggested that the evacuation may have been the consequence of fortuitous circumstances, including a magnitude 4.7 foreshock that possibly led people to leave their homes spontaneously.

The *Parkfield earthquake prediction experiment* was probably the most organized and well-documented earthquake prediction attempt so far. It was observed that the Parkfield segment of the San Andreas fault hosted earthquakes with magnitudes around 6 at rather regular intervals, namely in 1857, 1881, 1901, 1922, 1934, and 1966. Excluding the 1934 event, this gives a periodicity of 22 ± 4 years. A 95% chance of a major earthquake between 1988 and 1993 was predicted, but it did not happen. Eventually, a magnitude 6 event occurred near Parkfield on 28 September 2004, without any obvious precursor.

The occurrence of foreshocks, as in Haicheng, or a nearly regular seismicity pattern, as in Parkfield, are only two among many observations that might announce a larger earthquake. Others include unusual animal behavior, small changes in Earth structure (e.g., the ratio of P-wave and S-wave velocities, α/β), increased emissions of the gas radon (a radioactive decay product of uranium, contained in many crustal rocks), and electromagnetic anomalies.

Though many of these observations were once thought to be promising, none of them proved to be truly diagnostic. Only a tiny fraction of the many small earthquakes may be regarded as foreshocks of a major one; Earth structure may change over time as a consequence of fluid migration or volcanic processes; animals may just be sick or affected by the weather. Animals may in fact sense low-amplitude P-waves and act as an early warning of surface waves, which is often confused with prediction.

The failure of any current observations to predict earthquakes unambiguously – that is, without false alarms – severely reduces their practical usefulness. The large uncertainties inherent in earthquake prediction result in an additional social, political, and economic dimension. The difficulties are illustrated by the following scenario. Suppose that seismologists conclude that an earthquake with magnitude 7 or greater will take place sometime in a known month of a given year under a specific urban center. No more precise details are possible; in particular, the time of occurrence cannot be determined more exactly. Publication of this kind of prediction would cause great consternation, even panic, which for some of the population could be as disastrous as the earthquake itself. Should the warning be withheld? If the entire area, with its millions of inhabitants, were to be evacuated, the economic dislocation would be enormous. How long should the evacuation last, when every day is economically ruinous? Clearly, earthquake prediction is only useful in a practical sense when it is accurate in both place and time. The responsible authorities must also be provided with a reasonable estimate of the expected magnitude, from which the maximum intensity may be gauged. The problem then passes out of the domain of the scientist and into that of the politician. However, if the earthquake prediction is a failure, the consequences will rebound with certainty on the scientist, and the credibility of future predictions will be seriously reduced.

Following the boom of earthquake prediction research in the 1970s and 1980s, it has meanwhile become clear that truly useful predictions of major earthquakes are currently not possible. Some scientists suggest that earthquakes are inherently unpredictable, as is, for example, the weather in Paris exactly 10 years from now. Owing to the daunting difficulty of earthquake prediction, the modern research focus has shifted to forecasting and early warning.

7.3.1.2 Earthquake Forecasting and Probabilistic Seismic

Hazard Analysis

The main goal of earthquake forecasting is to provide engineers with information needed to build structures that withstand a given level of seismic ground shaking. For this, it aims to quantify the long-term probability that this level is exceeded within a predefined number of years. It is therefore also referred to as *probabilistic seismic hazard analysis* (PSHA).

In contrast to earthquake prediction, PSHA does not consider one, but *all* possible earthquakes, E_n, that may cause damage at a site in the following decades. The index $n = 1, 2, 3, \ldots$ labels different earthquakes. It may thus stand for a number of seismically active faults in the vicinity of the site, and for earthquakes with a range of magnitudes that may occur on these faults. In seismically active regions with many faults capable of producing large earthquakes, the total number of possible earthquakes (the maximum n) will be very large.

The engineer's interest is not in any one of the earthquakes, but in the ground shaking it may produce at the site. Therefore, it first needs to be defined what ground shaking is supposed to mean exactly. The most commonly used measures of ground shaking are the *peak ground acceleration* (PGA) and the *spectral acceleration* (SA) at a given oscillation period, normalized by the gravitational acceleration at the Earth's surface, $g = 9.81\ \mathrm{m\,s^{-2}}$. The period at which the spectral acceleration is considered depends on the longest resonance period of the building, which is the period at which ground shaking often produces maximum damage. As a rule of thumb, the longest resonance period of a building in seconds is roughly equal to the number of stories divided by 10. For example, the resonance period of a 100-story high-rise is about 10 s, and so it will respond most strongly to a 10 s seismic wave.

If one of the many possible earthquakes, E_n, has actually occurred in the past, and if it happened to be recorded at the site, ground motion is known for future recurrences of exactly that earthquake.

However, the vast majority of possible earthquakes have either never occurred before, or they have not been recorded in the right place. Therefore, a *ground motion model* (GMM), historically called an *attenuation relation* or *ground motion prediction equation*, is needed to anticipate the peak ground acceleration or spectral acceleration. Empirically, the logarithm of ground motion is almost linearly related to the magnitude M_n of the earthquake E_n and to the logarithm of its distance Δ_n from the site. (The magnitude can be any of the magnitudes defined in Section 7.2.6.2.) Thus, when ground shaking is displayed in terms of PGA, one approximately finds

$$\log(PGA_n) = f(M_n \Delta_n) = aM_n + b\log(\Delta_n + c) + d. \tag{7.12}$$

The constants *a, b, c*, and *d* that determine the function *f* must be estimated using actually observed earthquakes, as illustrated in Fig. 7.22. While Eq. (7.12) may indeed be used as a GMM, it ignores that ground motion also depends on many other factors, including the details of Earth structure between the fault and the site, local geology and soil properties, and the earthquake rupture process, which can be complicated, especially for larger events. As a consequence, actually observed values of log(PGA$_n$) scatter around the prediction of Eq. (7.12), thereby defining a *probability distribution* of ground motion, represented by the histogram in Fig. 7.22. For a given earthquake E_n with magnitude M_n and epicentral distance Δ_n, observed ground motion in terms of log(PGA$_n$) is more likely to be near the peak of the histogram, and less likely to be somewhere in its tails.

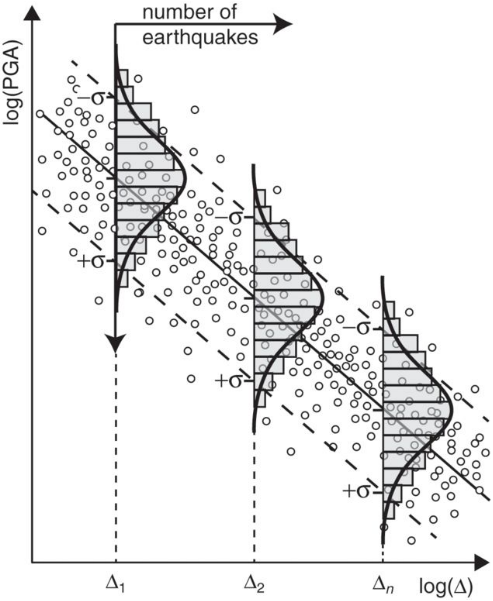


Fig. 7.22 Schematic illustration of a ground motion model (GMM). For a fixed magnitude, the logarithm of peak ground acceleration (PGA) for individual earthquakes (small circles) is empirically found to depend almost linearly on the logarithm of epicentral distance, Δ. For any epicentral distance, for example Δ_1, Δ_2, …, Δ_n, the scatter around the best-fitting line (black solid) can be displayed in the form of a histogram. The histogram can be approximated by a Gaussian curve (Eq. 7.13), the standard deviation of which is shown in the form of dashed lines.

Ground motion histograms as in Fig. 7.22 can often be approximated by a Gaussian distribution,

$$p_n = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \left[\log\left(PGA_n - f(M_n, \Delta_n) \right) \right]^2 / 2\sigma^2 \right). \tag{7.13}$$

The standard deviation σ describes how close the actually observed ground motion is to the prediction provided by the function *f* from Eq. (7.12). Large values of σ indicate wide scattering around $f(M_n, \Delta_n)$, and vice versa. The histogram or probability density in Fig. 7.22 provides a first means to compute the probability P_n of exceeding a critical ground motion level $\log(\mathrm{PGA}^{\mathrm{crit}})$ that may pose a serious risk to an engineering structure. In fact, P_n is equal to the sum of all histogram bars for which $\log(\mathrm{PGA}_n) \geq \log(\mathrm{PGA}^{\mathrm{crit}})$. This corresponds to the area under the Gaussian distribution between $\log(\mathrm{PGA}^{\mathrm{crit}})$ and ∞.

Each of the possible earthquakes E_n has a probability of actually occurring within a given time interval, for example, one year. Assuming, for simplicity, that earthquakes on a fault repeat with some approximate periodicity, this probability corresponds to the annual recurrence rate r_n. For small earthquakes that occur frequently, r_n can be estimated from actual observations. In the case of larger ones, the Gutenberg–Richter relationship (Eq. 7.11) may provide constraints. Knowing r_n, the annual probability of earthquake E_n exceeding the critical ground motion $\log(\mathrm{PGA}^{\mathrm{crit}})$ is given by

$$R_n = r_n P_n. \tag{7.14}$$

Summing over all possible earthquakes finally gives the total probability R_{tot} of exceeding the critical ground motion within one year:

$$R_{\mathrm{tot}} = \sum_n R_n. \tag{7.15}$$

The concept of a probabilistic seismic hazard analysis can be illustrated with the oversimplified example of only two potential earthquakes, E_1 and E_2. While E_1 with magnitude $M_1 = 6$ recurs every five years ($r_1 = 0.2\ \mathrm{year}^{-1}$), E_2 with magnitude $M_2 = 7$ recurs every 50 years ($r_2 = 0.02\ \mathrm{year}^{-1}$). Both are assumed to be located at 10 km distance from the site of interest ($\Delta_1 = \Delta_2 = 10$ km). A GMM based on seismic observations in the western USA relates M_n and Δ_n to the natural logarithm of relative PGA at the site (Cornell et al., 1979):

$$\ln(PGA_n) = 0.859M_n - 1.803\ln(\Delta_n + 25.0) - 0.152. \quad (7.16)$$

Equation (7.16) is a special form of Eq. (7.12), where the coefficients have been estimated as $a = 0.859$, $b = -1.803$, $c = 25.0$, and $d = -0.152$. The GMM does not provide a perfect prediction of ground motion, and actual PGA observations will scatter around the values provided by Eq. (7.16). This scatter defines a Gaussian distribution of PGA, as in Eq. (7.13). The Gaussian curves for the earthquakes E_1 and E_2 are displayed in Fig. 7.23. The individual probabilities $P_1 = 0.12$ and $P_2 = 0.63$ to exceed a critical PGA of 0.5 g are equal to the shaded areas under the Gaussians. Using Eqs. (7.14) and (7.15) gives a total annual exceedance probability of $R_{tot} = 0.036$.

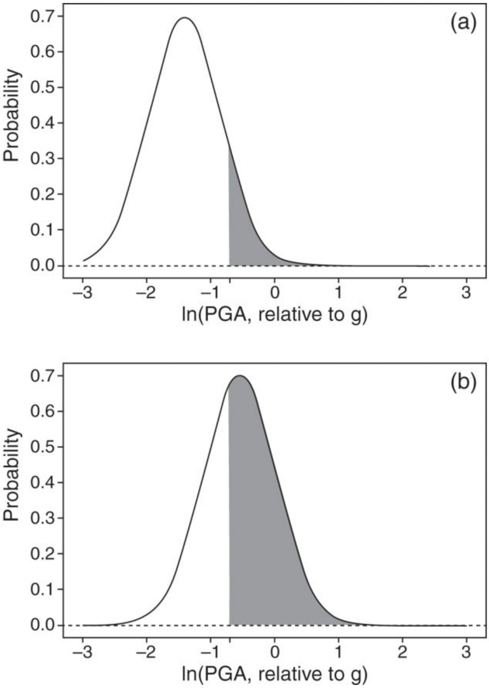


Fig. 7.23 Gaussian probability distributions for ln(PGA) relative to g for a magnitude 6 earthquake (a) and a magnitude 7 earthquake (b). The peak values are the ln(PGA) values predicted by the GMM in Eq. (7.16) for the western USA. The standard deviation, which indicates the scatter of actual observations around the prediction, is

σ = 0.57. The shaded area under the curves is equal to the probability to exceed a critical PGA of 0.5 g (ln(0.5) ≈ –0.693).

Realistic seismic hazard analyses are of course considerably more involved. Modern GMMs may account for local geology or the details of rupture propagation on a complicated fault, and the number of potential earthquakes is enormous. The 2013 European Seismic Hazard Map (Giardini et al., 2014) shown in Fig. 7.24 is based on more than 1200 mapped faults with a total length of about 68,000 km. The map displays the PGA that is likely to be exceeded with a 10% probability within 50 years.

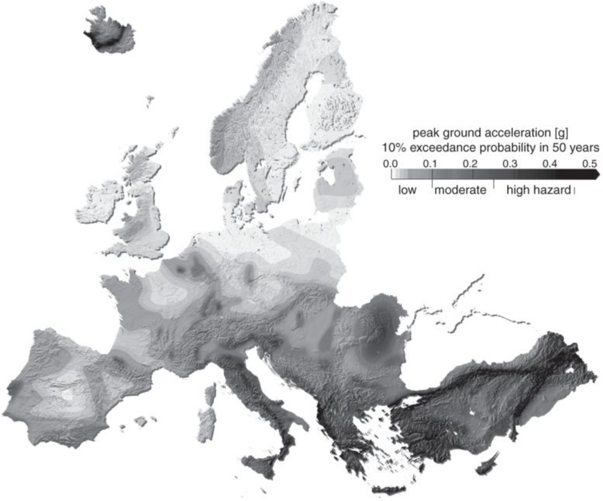

peak ground acceleration [g]
10% exceedance probability in 50 years
0.0
0.1
0.2
0.3
0.4
0.5
low
moderate
high hazard

Fig. 7.24 The 2013 European Seismic Hazard Map, displaying the PGA that is likely to be exceeded with a 10% probability during a period of 50 years. The dark streak in northern Turkey, for example, traces the North Anatolian fault, where destructive earthquakes with magnitudes above 7 occur roughly every 10 years. (Courtesy of Laurentiu Danciu and Domenico Giardini, based on work by Giardini et al., 2014).

Though the construction of modern seismic hazard maps is a complex procedure, they still rely on numerous simplifications and assumptions. For example, earthquakes may not simply recur in some nearly periodic fashion, especially far from plate boundaries where there is no continuous strain accumulation. Furthermore, the parameterization of the GMMs (e.g., Eqs. 7.12 and 7.16) is largely subjective. Many other empirical equations have been proposed, and they may provide different results for magnitudes and distances for which few observations have been made.

Unfortunately, relatively low seismic hazard does not mean that large seismic ground motion is entirely impossible. Seismic hazard is quantified in terms of probabilities, and even a very unlikely event still has some small probability to actually happen. A recent example is the magnitude 9.1 Tohoku earthquake that struck Japan on March 11, 2011. Its epicenter was located in a region of relatively low seismic hazard.

While an elevated seismic hazard indicates the potential of large ground motion, it need not imply large destruction. In fact, damage to buildings and infrastructure strongly depends on the quality of the construction. The study of the related socioeconomic impact, which is beyond the scope of this book, is called *seismic risk analysis*.

7.3.1.3 Earthquake Early Warning

In contrast to prediction and forecasting, early warning becomes effective immediately after the occurrence of an earthquake or even while the fault is still rupturing. The goal of early warning is to enable short-term measures to protect people, buildings, and infrastructure. These measures may include the closing of bridges and tunnels, the shut down of nuclear power plants, the stopping of high-speed trains, and the evacuation of buildings.

Though being technically involved, the basic concept of early warning is simple. It primarily rests on the detection of fast-traveling body waves near the epicenter and the rapid estimation of earthquake location and magnitude. Based on these estimates, a warning may be issued before the arrival of the slow, high-amplitude surface waves. Today, early warning systems are operational in regions where seismic hazard is particularly high. This includes Japan, Taiwan, and parts of

California and British Columbia.

Since the propagation speed of seismic surface waves is around 2–4 km s^{-1}, the time before their arrival is only on the order of several seconds or tens of seconds. In the case of tsunami waves, which only travel at a speed of around 0.2 km s^{-1} (Section 7.3.2), the warning time can be much longer, potentially allowing people to reach places with higher elevation.

7.3.2 Secondary Effects of Earthquakes: Landslides, Tsunamis, Fires, and Fatalities

In addition to ground shaking, earthquakes are frequently accompanied by secondary effects: landslides, seismic sea waves, and conflagrations. These rather exceptional effects cannot be included conveniently in the definition of earthquake *intensity*, because their consequences cannot be easily generalized or quantified. For example, once a major fire has been initiated, other factors not related directly to the size of the earthquake (such as aridity of foliage, combustibility of building materials, availability and efficiency of fire-fighting equipment) determine how it is extinguished.

A major hazard associated with large earthquakes in mountainous areas is the activation of large landslides, which can cause destruction far from the epicenter. In 1970, an earthquake in Peru with magnitude 7.8 and shallow focal depth of 40 km caused widespread damage and a total death toll of about 66,000. High in the Cordillera Blanca mountains above the town of Yungay, about 15 km away, an enormous slide of rock and ice was released by the tremors. Geologists later speculated that a kind of “air cushion” had been trapped under the mass of rock and mud, enabling it to acquire an estimated speed of 300–400 km h^{-1}, so that it reached Yungay less than five minutes later. Over 90% of the town was buried under the mud and rock, in places to a depth of 14 m, and 20,000 lives were lost.

When a large earthquake occurs under the ocean, it can activate a seismic sea wave known as a *tsunami*, which in Japanese means “harbor wave.” This special type of sea wave (Box 7.1) is a long-wavelength disturbance of the ocean surface that can be triggered by large-scale collapse or uplift of part of the ocean floor, by an underwater landslide, or by submarine volcanism. The magnitude 9.0 Sumatra–Andaman Islands earthquake on December 26, 2004, had a rupture zone more than 1200 km in length and 150 km in width, with displacements on the fault of 20 m (Lay et al., 2005). This probably resulted in uplift of the ocean bottom by several meters, which initiated a disastrous tsunami in the Indian Ocean.

Box 7.1 Tsunami

When the ocean floor is abruptly lifted or dropped by a submarine earthquake or landslide, the entire water column is pushed up and down. More than 95% of the potential energy of the displaced water is gravitational and less than 5% is elastic energy resulting from compression of the ocean floor or water column. The potential energy of the vertical motion is converted to kinetic energy and propagates away from the source as a tsunami. The propagation of ocean waves is a complex nonlinear problem that cannot be handled here, but approximate solutions are useful for the special case of a tsunami.

Ocean surface waves with periods shorter than 50 s are confined to the top kilometer of the ocean and cannot be excited by motions of the sea floor. The propagation of ocean waves is in general dispersive, i.e., the wave velocity depends on its wavelength and period. Figure B7.1.1 shows the computed wave velocities for different periods in water of different depths. The wave speeds are dispersive for periods shorter than about 300 s (5 minutes). These waves have very long wavelengths, much greater than the ocean depth.

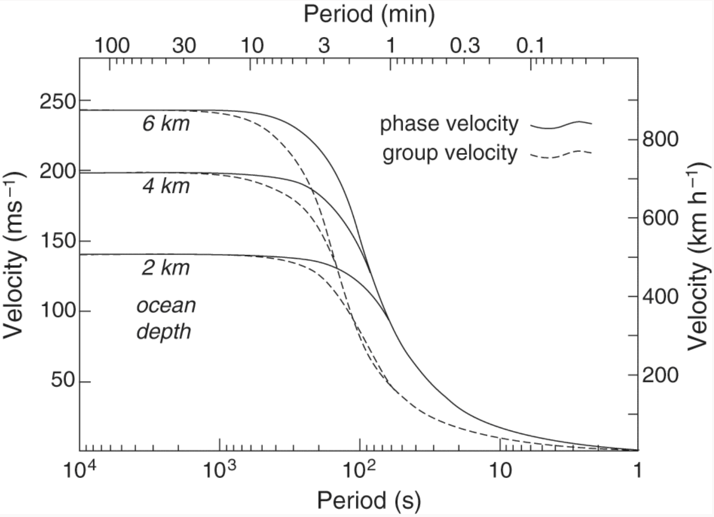


Fig. B7.1.1 Dependence of the phase velocity (c) and group-velocity (U) of a tsunami on the period of the wave for different ocean depths.

Data source: S.N. Ward, personal communication, 2006.

For a wave with period T (angular velocity $\omega = 2\pi/T$) and wavelength λ (wave number $k = 2\pi/\lambda$), the dispersion in an ocean of depth d is governed by the relation

$$\omega^2 = gk\tanh(kd), \tag{1}$$

where g is the acceleration of gravity and tanh(x) is the hyperbolic tangent function of x:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \begin{cases} x, & \text{if } x \text{ small} \\ 1, & \text{if } x \text{ large} \end{cases}. \tag{2}$$

If the wavelength λ is very much greater than the ocean depth, tanh(kd) can be replaced in Eq. (1) by (kd) and the dispersion relation becomes

$$\omega^2 = gk(kd) = k^2 gd$$
$$\omega = k\sqrt{gd}.$$
(3)

Analogously to surface waves, a tsunami travels across an ocean as a packet of waves of different period. The *phase velocity* (c) of a wave is the speed of an individual phase in the packet. Using Eq. (3) and Eq. (6.24), the phase velocity for a tsunami is given by

$$c = \frac{\omega}{k} = \sqrt{gd}.$$
(4)

The *group velocity* (U) is the propagation speed of the envelope of a wave-packet, and thus is the speed with which the energy of the tsunami travels. Using Eq. (3) and Eq. (6.44), the group velocity for a tsunami is given by

$$U = \frac{\partial \omega}{\partial k} = \sqrt{gd}.$$
(5)

For water depths that are much less than the wavelength, the phase and group velocities are equal and the propagation of a tsunami is non-dispersive. These relations are valid for wavelengths greater than about 11 times the water depth. Over an ocean basin that is 4 km deep the propagation velocity is 200 m s^{-1} (720 km h^{-1}) and the wavelengths for periods of

1000–2000 s (around 15–30 min) are 200–400 km. When the tsunami approaches shore and the water depth decreases, its velocity slows. The advancing wave loses kinetic energy, part of which is lost as friction with the ocean bottom and part converted into potential energy. This is evident as increase in wave height, from a few tens of centimeters over the open ocean to many meters close to land. The contact with land is less often as a ferocious breaking wave than as a rapid increase in sea level – as during high tides – accompanied by fierce currents which may sweep far inland and withdraw in the same way.

A tsunami propagates throughout an ocean basin as a wave with period T of around 15–30 min. The entire water column participates in the motion. As a result, the velocity of the wave, V, depends on the water depth, d, and the acceleration of gravity, g, and is given by:

$$V = \sqrt{gd}. \tag{7.17}$$

Over an ocean basin with water depth greater than 4 km, the tsunami velocity is higher than 200 m s^{-1}, and the wavelength (equal to the product VT) may measure 200 km (Box 7.1). The amplitude of a tsunami over the open ocean is comparatively small; the Sumatra tsunami measured about 80–100 cm from crest to trough in the open Indian Ocean. Despite the speed of the wave, an observer on a ship would be scarcely aware of the passage of such a low-amplitude, long-wavelength disturbance. However, on approaching shallower water the leading part of the tsunami slows down and tends to be overridden by the following water mass, so that the height of the wave increases. The wave height may be amplified to several meters by the shapes of the sea bottom and the coastline.

In 1896 a large earthquake raised the ocean bottom off the southern shore of Japan and initiated a tsunami that raced ashore with an estimated wave height of more than 20 m, causing 26,000 fatalities. One of the best-studied tsunamis was set off by a great earthquake in the Aleutian Islands in 1946. It traveled across the Pacific and several hours later reached Hilo, Hawaii, where it swept ashore and up river estuaries as a wave 7 m high. The 2004 Sumatra earthquake triggered the worst tsunami in recorded history, with probably more than 250,000 fatalities in Indonesia, Thailand, Sri Lanka, and other countries bordering the Indian Ocean, as far away as Somalia. On March 11, 2011, the magnitude 9.1 Tohoku earthquake caused a tsunami with a maximum height of 40 m, reported to have traveled up to 10 km inland in some places. The tsunami caused nuclear meltdowns in three reactors at the Fukushima power plant.

When a major earthquake is detected that can produce a tsunami, a warning can be issued to alert endangered regions to the imminent threat. The system works well far from the source, where warning times are sufficiently long. It may take several hours for the tsunami to arrive at a distant location, as illustrated by records of tsunami propagation for the 1960 Chilean and 1964 Alaskan earthquakes (Fig. 7.25). This allows time to warn and evacuate the population in many places far from the epicenter. However, tsunami casualties may still occur near the epicenter of the earthquake.

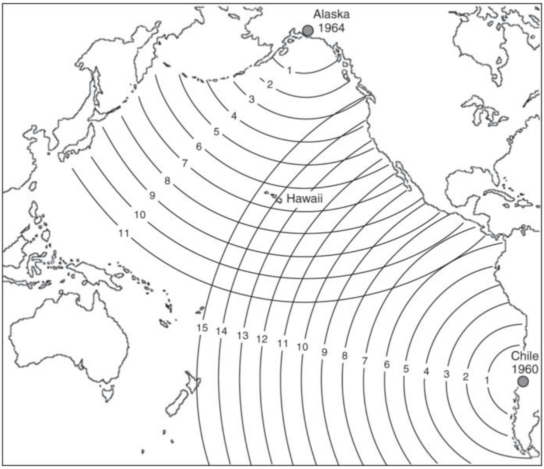


Fig. 7.25 Schematic illustration of the propagation of tsunami waves across the Pacific Ocean following the 1960 Chilean earthquake ($M_\mathrm{w} = 9.5$) and the 1964 Alaskan earthquake ($M_\mathrm{w} = 9.2$). Numbers on the wavefronts show the travel times in hours. Due to variations in ocean depth, actual tsunami wavefronts are not perfectly circular.

Source: http://pubs.usgs.gov/publications/text/tsunamis.html.

In addition to causing direct damage to man-made structures, an earthquake can disrupt subterranean supply routes (e.g., telephone, electrical, and gas lines), which in turn increases the danger of explosion and fire. Aqueducts and underground water pipelines may be broken, with serious consequences for the inhibition or suppression of fires. The San Francisco earthquake of 1906 was very powerful. The initial shock caused widespread damage, including the disruption of water supply lines. But a great fire followed the earthquake, and because the water supply lines were broken by the tremor, it could not be extinguished. The greatest damage in San Francisco resulted from this conflagration.

7.3.3 Induced Seismicity

In March 1962 the US Army began to dispose of liquid toxic waste from the manufacture of chemical weapons at Rocky Mountain Arsenal by injection into a well, more than 3 km deep, near Denver, Colorado. Although the region had been nearly devoid of earthquake activity for the preceding 80 years, earthquakes began to occur several weeks after pumping started. Until 1965, when waste injection was halted, more than 1000 earthquakes were recorded. They were mostly very small, but some had magnitudes as high as 4.8. By 1967 induced earthquakes had migrated as far as 10 km away from the injection site. Though injection terminated in February 1966, elevated seismic activity continued for about two decades. A magnitude 4.3 earthquake occurred in April 1981.

It has long been known that earthquakes may be induced not only by the injection of fluids into deep rock formations, but also by surface and underground mining, and by the withdrawal of oil and gas from reservoirs. While global seismicity is clearly focused along tectonic plate boundaries (Fig. 7.5), considerable potential for earthquakes also exists within continental interiors, despite the comparatively low deformation rates. The main reason is that shear stress levels within tectonic plates are commonly found to be very close to the actual strength of crustal rocks. Therefore, even small perturbations of the local stress regime may cause significant earthquakes.

Damaging earthquakes related to human activity received considerable attention in recent years due to their dramatically increasing number. For example, more than 300 earthquakes with magnitudes above 3 occurred in the central and western USA in the three years between 2010 and 2013. For comparison, the average number of $M > 3$ earthquakes per year between 1967 and 2000 was about 20. This translates into an increase by a factor of around 5.

Induced seismicity is a remarkably diverse phenomenon, which makes management of the related seismic risk a very challenging task. Induced earthquakes may occur near the source of the stress perturbation, or deep below and many kilometers away. Far from the source, induced earthquakes release stress on preexisting faults that may not have been known before. Induced earthquakes can occur shortly after the industrial activity begins, but also long after it has been underway or ceased.

Among the many causes of induced seismicity, the injection of fluids into underground formations plays a particularly important role. Fluids may be injected under high pressure for the purpose of *hydraulic fracturing*, colloquially known as *fracking*. The high-pressure fluids create fractures within rocks, thereby increasing their permeability. As a result, oil and gas may be recovered from otherwise unproductive formations, or hot fluids may be circulated to enhance geothermal energy production. The extraction of oil and gas, with or without hydraulic fracturing, typically produces large volumes of waste water – that is, highly saline or even toxic water that cannot be used for irrigation, drinking, or any other useful purpose. To avoid environmental contamination, waste water is often pumped back into deep underground formations where it may migrate along pathways that are difficult to know beforehand. The magnitude 5.7 event that occurred on November 6, 2011 in central Oklahoma is often considered to be among the largest induced earthquakes. It was most likely related to the injection of waste water into a depleted oil reservoir.

Induced seismicity related to the extraction of hydrocarbons can be exemplified by the case of the Groningen gas field, located in the northeastern Netherlands. This is the largest gas field of Western Europe, around 2100 billion m^3 of gas having been produced between 1963 and 2012. (This volume corresponds to a cube of 13 km edge length.) The region is tectonically stable, without any record of seismicity before the occurrence of the first induced earthquakes in 1991 – that is, 28 years after the beginning of gas extraction. The largest event, with magnitude 3.6, was observed in 2012. While the magnitude was relatively low, seismic intensities as high as VI (slight damage to houses and infrastructure, see Table 7.1) resulted from the shallow depth of the event (3 km) and the very soft soil in the region. Induced seismicity in the Groningen field is closely linked to compaction of the reservoir caused by the extraction of large volumes of gas. Compaction leads to stress perturbations, which can trigger earthquakes on some of the more than 1800 faults that have been identified in the area by seismic imaging. Compaction also leads to subsidence, reaching as much as 30 cm in parts of the reservoir.

While the long-term seismic hazard related to industrial activities is to some extent controllable, for example, by lowering waste water injection rates, it currently seems unlikely that the occurrence of large man-made earthquakes can be entirely avoided, even long after activities have ceased.

7.3.4 Nuclear Explosion Monitoring

Since 1963 most nuclear explosions have been conducted underground to prevent the dangerous radioactive fallout that accompanies nuclear explosions conducted under water or in the atmosphere. The detection and monitoring of such testing activity became important tasks for seismologists. The bilateral Threshold Test Ban Treaty of 1974 between the former Soviet Union and the USA prohibited underground testing of nuclear devices with a yield greater than 150 kilotons of TNT equivalent, which corresponds roughly to an earthquake with magnitude about 6.

On September 10, 1996, the Comprehensive Nuclear-Test-Ban Treaty (CTBT) was adopted by the United Nations General Assembly. The CTBT bans all nuclear explosions for any purpose, military or civilian. Since a small number of countries has not yet signed or ratified the treaty, it is formally not yet effective. Nevertheless, a system to monitor compliance with the CTBT is already in place. In addition to seismology, it is based on measurements of acoustic waves and radionuclides in the atmosphere. The seismological detection of nuclear explosions at distances of several thousand kilometers, and their discrimination from an enormous number of small earthquakes, poses a monumental challenge to seismologists.

In order to achieve the high detection capability needed to monitor underground testing, many so-called seismic arrays have been set up. An array consists of several individual seismometers, with spacing on the order of a kilometer or less, that feed their output signals into a centralized data-processing center. By filtering, delaying, and summing the signals of the individual instruments, the incoherent noise is reduced and the coherent signal is increased, thus improving the signal-to-noise ratio significantly over that for a single sensor. A local seismic wave arrives at the array on an almost horizontal path and triggers the individual seismometers at successively different times, whereas a teleseismic arrival from a very distant source reaches all seismometers in the array at nearly the same time along a steeply inclined path. The development of seismic arrays permitted the analysis of distant weak events. The enhanced sensitivity led to several advances in seismology. Features of the deep structure of the Earth (e.g., the inner core) could be investigated, earthquake location became more accurate, and the analysis of focal mechanisms received a necessary impetus. These improvements were essential because of the need to identify correctly the features that distinguish underground nuclear explosions from small earthquakes.

An earthquake is the result of sudden motion of crustal blocks on opposite sides of a fault plane. The radiation pattern of P-wave amplitude has four lobes of alternating compression and dilatation (Fig. 7.26). The first motions at the surface of the Earth are either pushes away from the source or tugs toward it, depending on the geometry of the focal mechanism. In contrast, an underground explosion causes outward pressure around the source. The first motions at the surface are all pushes away from the source. Hence, focal mechanism analysis provides an important clue to the nature of the recorded event. Moreover, an explosion produces predominantly P-waves, while earthquakes are much more efficient in also generating surface waves. Consequently, the relative amplitudes of the long-period surface wave part of the record to the short-period P-wave part are much higher for an earthquake than for an explosion (Fig. 7.26).

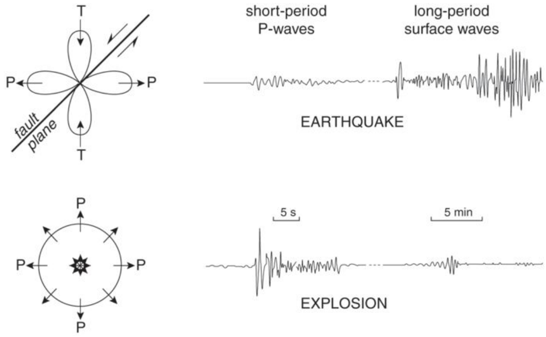


Fig. 7.26 Comparison of P-wave radiation patterns and the relative amplitudes of long-period and short-period surface waves for an

earthquake and a nuclear explosion.

Source: P.G. Richards, Seismic monitoring of nuclear explosions, in: *The Encyclopedia of Solid Earth Geophysics*, D.E. James, ed., pp. 1071–1089, New York: Van Nostrand Reinhold, 1989.

Further discrimination criteria are the epicentral location and the focal depth. Intraplate earthquakes are much less common than earthquakes at active plate margins, so an intraplate event might be suspected to be an explosion. If the depth of a suspicious event is determined with a high degree of confidence to be greater than about 15 km, one can virtually exclude that it is an explosion. Deeper holes have not been drilled due to the great technical difficulty, e.g., in dealing with the high temperatures at such depths.

7.4 Internal Structure of the Earth

7.4.1 Introduction

It is well known that the Earth has a molten outer core. What is now general knowledge was slow to develop. In order to explain the existence of volcanoes, some nineteenth-century scientists postulated that the Earth must consist of a rigid outer crust around a molten interior. It was also known in the last century that the mean density of the Earth is about 5.5 times that of water. This is much larger than the known specific density of surface rocks, which is about 2.5–3. From this it was inferred that density increases toward the Earth's center under the effect of gravitational pressure. The density at the Earth's center was estimated to be comparatively high, greater than 7000 kg m^{-3} and probably in the range 10,000–12,000 kg m^{-3}. It was known that some meteorites had a rock-like composition, while others were much denser, composed largely of iron. In 1897 Emil Wiechert, who subsequently became a renowned German seismologist, suggested that the interior of the Earth might consist of a dense metallic core, cloaked in a rocky outer cover. He called this cloak the "Mantel," which later became anglicized to mantle.

The key to modern understanding of the interior of the Earth – its density, pressure, and elasticity – was provided by the invention of the Milne seismograph. The progressive refinement of this instrument and its systematic employment worldwide led to the rapid development of the modern science of seismology. Important results were obtained early in the twentieth century. The Earth's fluid core was first detected seismologically in 1906 by Richard D. Oldham. He observed that, if the travel times of P-waves observed at epicentral distances of less than 100° were extrapolated to greater distances, the expected travel times were less than those observed. This meant that the P-waves arriving at large epicentral distances were delayed in their passage through the Earth. Oldham inferred from this the existence of a central core in which the P-wave velocity was reduced. He predicted that there would be a region of epicentral distances (a "shadow zone") in which P-waves could not arrive. About this time it was found that

P- and S-waves passed through the mantle but that no S-waves arrived beyond an epicentral distance of 105°. In 1914 Beno Gutenberg verified the existence of a shadow zone for P-waves in the range of epicentral distances between 105° and 143°. Gutenberg also located the depth of the core–mantle boundary with impressive accuracy at about 2900 km. A modern estimate of the radius of the core is 3485 ± 3 km, giving a mantle 2885 km thick. Gutenberg also predicted that P-waves and S-waves would be reflected from the core–mantle boundary. These waves, known today as PcP and ScS waves, were not observed until many years later. In honor of Gutenberg the core–mantle boundary is sometimes referred to as the *Gutenberg seismic discontinuity*.

While studying the P-wave arrivals from an earthquake in Croatia in 1909, Andrija Mohorovičić found only a single arrival at distances close to the epicenter (the P_g phase, where "g" indicates that the wave travels through the upper, largely "granitic" crust). Beyond about 200 km there were two arrivals; the P_g phase was overtaken by another arrival (the P_n phase, where "n" stands for a wave traveling through the "normal" mantle), which had evidently traveled at higher speed. Mohorovičić identified P_g as the direct wave from the earthquake and P_n as a refracted wave (equivalent to a head wave) that traveled partly in the upper mantle. Mohorovičić calculated velocities of 5.6 km s_{-1} for P_g and 7.9 km s_{-1} for P_n, and estimated that the sudden velocity increase occurred at a depth of about 54 km. This seismic discontinuity is now called the *Mohorovičić discontinuity*, or *Moho* for short. It represents the boundary between the crust and mantle. The crustal thickness is very variable. It averages about 33 km, but measures as little as 5 km under oceans and as much as 60–80 km under some mountain ranges.

The depth of the Moho inferred from seismic data, e.g., reflected waves observed at the surface, is commonly referred to as the *seismological Moho*. It represents an estimate of the *petrological Moho*, that is, the actual depth where rock type changes.

It is now known that the crust is not homogeneous but has a vertically layered structure. In 1925 Victor Conrad separated arrivals from a Tauern (Eastern Alps) earthquake of 1923 into P_g and S_g waves in an upper crustal layer and faster P* and S* waves that traveled with velocities 6.29 km s^{-1} and 3.57 km s^{-1}, respectively, in a deeper layer. (Alternatively, P* and S* are termed P_b and S_b, where “b” stands for “basaltic”.) Because the P* and S* velocities are significantly slower than corresponding upper mantle velocities, Conrad deduced that they were head waves from a lower crustal layer. The interface separating the continental crust into an upper and a lower layer is called the *Conrad discontinuity*. Influenced by early petrological models of crustal composition and by comparison with seismic velocities in known materials, seismologists referred to the upper and lower crustal layers as the granitic layer and the basaltic layer, respectively. This petrological separation is now known to be overly simplistic. In contrast to the Moho, which is a sharp discontinuity nearly everywhere, the Conrad discontinuity is poorly defined or absent in some areas.

The core shadow zone and its interpretation in terms of a fluid core were well established in 1936 when Inge Lehmann, a Danish seismologist, reported weak P-wave arrivals within the shadow zone. She interpreted these in terms of an inner core with higher seismic velocity. However, the existence of the inner core remained controversial for many years. Improved seismometer design, digital signal treatment, and the setting up of seismic arrays have provided corroborating evidence. The existence of a solid inner core is also supported by analyses of the Earth’s normal modes.

The gross internal structure of the Earth is modeled as a set of concentric shells, corresponding to the inner core, outer core, and mantle (Fig. 7.27). An important step in understanding this layered structure has been the development of travel time curves for seismic rays that passed through the different shells. To facilitate identification of the arrivals of these rays on seismograms, a convenient shorthand notation is used. A P- or S-wave that travels from an earthquake directly to a seismometer is labeled with the appropriate letter P or S; until the margin of the core shadow zone, the P- and S-waves follow slightly different curved paths because the increase with depth of P-wave and S-wave velocities is not identical. (The curvature, as explained in Section 6.3.6.3, arises from the increase of seismic velocity with depth.) A wave that reaches the seismometer after being reflected once from the surface is labeled PP (or SS), as its path consists of two identical P- or S-segments.

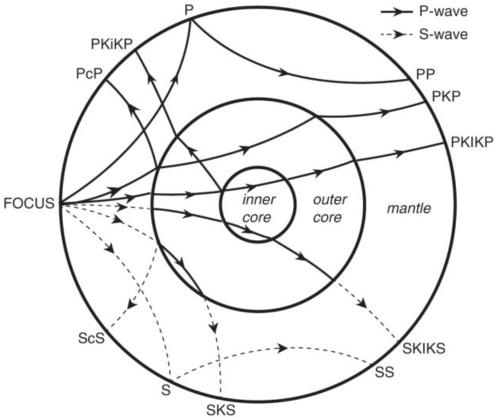


Fig. 7.27 Seismic ray paths of some important refracted and reflected P-wave and S-wave phases from an earthquake with focus at the Earth’s surface.

Source: B. Gutenberg, *Physics of the Earth’s Interior*, 240 pp., New York: Academic Press, 1959.

The energy of an incident P- or S-wave is partitioned at an interface into reflected and refracted P- and S-waves (see Section 6.3.5.3). A P-wave incident on the boundary between mantle and fluid outer core is refracted toward the normal to the interface, because the P-wave velocity drops from about 13 km s^{-1} to about 8 km s^{-1} at the boundary. After a second refraction it emerges beyond the shadow zone and is called a PKP wave (the letter K stands for Kern, the German word for core). An S-wave incident at the same point has a lower mantle velocity of about 7 km s^{-1}. Part of the incident energy is converted to a P-wave in the outer core, which has a higher velocity of 8 km s^{-1}. The refraction is away from the normal to the interface. After a further refraction the incident S-wave reaches the surface as an SKS phase. A P-wave that travels through mantle, fluid core, and inner core is labeled PKIKP. Each of these rays is refracted at an internal interface. To indicate seismic phases that are reflected at the outer core boundary the letter c is used, giving rise, for example, to PcP and ScS phases (Fig. 7.28). Reflections from the inner core are designated with the letter i, as for example in the phase PKiKP.

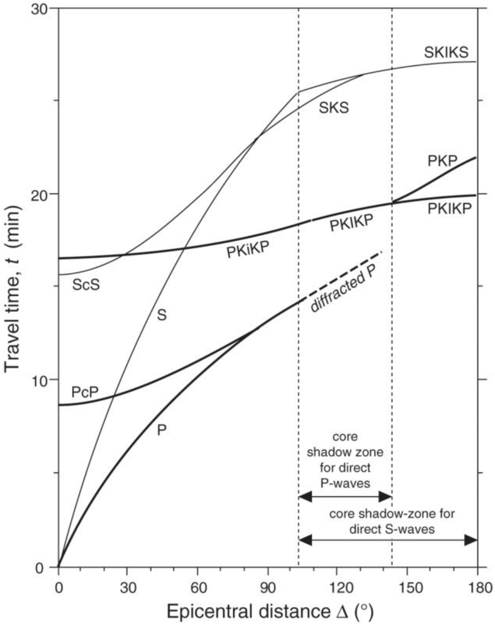


Fig. 7.28 Travel time versus epicentral distance ($t - \Delta$) curves for some important seismic phases.

Data source: H. Jeffreys and K.E. Bullen, *Seismological Tables*, 50 pp., London: British Association, Gray-Milne Trust, 1940.

7.4.2 Refractions and Reflections in the Earth's Interior

If it possesses sufficient energy, a seismic wave may be refracted and reflected – or converted from a P-wave to an S-wave, or vice versa – many times at the several seismic discontinuities within the Earth and at its free surface. As a result, the seismogram of a large earthquake contains numerous overlapping seismic signals, making the identification of individual phases a difficult task. Late-arriving phases that have been multiply reflected or that have traveled through several regions of the Earth's interior may be difficult to distinguish from earlier arrivals. In the period 1932–1939, Sir Harold Jeffreys and Keith E. Bullen analyzed a large number of good records of earthquakes registered at a worldwide, though sparse, distribution of seismic stations. In 1940 they published a set of tables giving the travel times of P- and S-waves through the Earth. A slightly different set of tables was reported by Beno Gutenberg and Charles F. Richter. The good agreement of the independent analyses confirmed the reliability of the results. The Jeffreys–Bullen seismological tables were used by the international seismic community as the standard of reference for many years.

The travel time of a seismic wave to a given epicentral distance is affected by the focal depth of the earthquake, which may be as much as several hundred kilometers. The travel time versus distance curves of some important phases are shown in Fig. 7.28 for an earthquake occurring at the Earth's surface. The model assumes that the Earth is spherically symmetric, with the same vertical structure underneath each place on the surface. This assumption works fairly well, although it is not quite true. Lateral variations of seismic velocity have been found at many depths within the Earth. For example, there are lateral differences in seismic velocity between oceanic and continental crust, and between oceanic and continental lithosphere. At even greater depths significant lateral departures from the spherical model have been detected. These discrepancies form the basis of the branch of seismology called *seismic tomography*, which we will examine later.

7.4.2.1 Seismic Rays in a Uniformly Layered Earth

It is important to understand clearly the relationship between the travel time (t) versus epicentral distance (Δ) curves and the paths of seismic waves in the Earth, like those shown in Fig. 7.27. Consider first an Earth that consists of two concentric shells representing the mantle and core (Fig. 7.29a). The P-wave velocity in each shell is constant, and is faster in the mantle than in the core ($\alpha_1 > \alpha_2$). The figure shows the paths of 18 rays that leave a surface source at angular intervals of 5°. Rays 1–12 travel directly through the mantle as P-waves and emerge at progressively greater epicentral distances. The convex upwards shape of the t–Δ curve is here due to the curved outer surface, the layer velocity being constant. Ray 13 is partially refracted into the core, and partially reflected (not shown in the figure). Because $\alpha_2 < \alpha_1$ the refracted ray is bent toward the Earth's radius, which is normal to the refracting interface. This ray is further bent on leaving the core and reaches the Earth's surface as a PKP phase at an epicentral distance greater than 180°. Rays 14 and 15 impinge more directly on the core and are refracted less severely; their epicentral distances become successively smaller and their travel times become shorter than ray 13, as indicated by branch 13–15 of the t–Δ curve. This branch is offset in time from the extrapolation of branch 1–12 because of the lower velocity in the core. The paths of rays 16, 17, and 18 (which is a straight line through mantle and core and emerges at an epicentral distance of 180°) become progressively longer, and a second branch 16–18 develops on the t–Δ curve. The two branches meet at a sharp point, or cusp. No P-waves reach the surface in the gap between rays 12 and 15 in this simple model. There is a *shadow zone* between the last P-wave that just touches the core and the PKP wave with the smallest epicentral distance. The existence of a shadow zone for P-waves is evidence for a core with lower P-wave velocities than the mantle. S-waves in the mantle follow similar ray paths, controlled by the distribution of S-wave velocity β. However, no direct S-waves arrive in the shadow zone, which indicates that the core must be fluid.

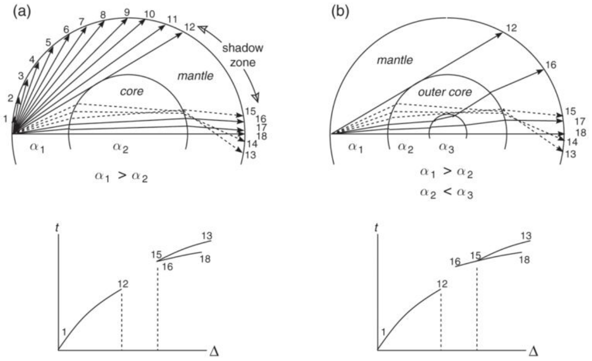

(a)
1
2
3
4
5
6
7
8
9
10
11
12
shadow zone
mantle
core
15
16
17
18
14
13
α_1
α_2
$\alpha_1 > \alpha_2$
t
1
12
15
16
13
18
Δ
(b)
mantle
outer core
12
16
15
17
18
14
13
α_1
α_2
α_3
$\alpha_1 > \alpha_2$
$\alpha_2 < \alpha_3$
t
1
12
16
15
13
18
Δ

Fig. 7.29 Seismic wave paths and their $t-\Delta$ curves for P-waves passing through a spherical Earth with constant velocities in the mantle, outer core, and inner core. (a) Development of a shadow zone when the mantle velocity (α_1) is higher than the outer core velocity (α_2). (b) Penetration of the shadow zone by rays refracted in an inner core with higher velocity than the outer core ($\alpha_3 > \alpha_2$).

Now suppose an inner core with a constant velocity α_3 that is higher than the velocity α_2 in the outer core (Fig. 7.29b). The paths of rays 1–15 are the same as before, through the mantle and outer core. The segments 1–12 and 13–15 of the t–Δ curve are the same as previously. Ray 16 impinges on the inner core and is sharply refracted away from the Earth's radius; on returning to the outer core it is again refracted, back toward the radius. After further refraction at the core–mantle interface this ray emerges at the Earth's surface at a smaller epicentral distance than ray 15, within the P-wave shadow zone, as a PKIKP arrival. Successive PKIKP rays are bent less strongly. The PKIKP rays map out a new branch 16–18 of the t–Δ curve (see also the travel time curves in Fig. 7.28).

7.4.2.2 Travel Time Curves for P-, PKP-, and PKIKP-Waves in the Earth

Though seismic velocities are not constant within the individual layers, the analysis from the previous paragraph helps to understand the ray paths of actual P, PKP, and PKIKP phases shown in Fig. 7.30. There is a shadow zone for direct P-waves between about 103° and 143°, and no direct S-waves are found beyond 103°. The shallowest PKP ray (A in Fig. 7.30) is deviated the furthest, emerging at an epicentral distance greater than 180°. Successively deeper PKP rays (B–E) emerge at ever-smaller epicentral distances until about 143°, after which the epicentral distance increases again to almost 170° (rays F, G). It was long believed that the boundary between inner and outer core was a transitional region with higher P-wave velocity, and that PKP rays traversing this region would again emerge at smaller epicentral distances (ray H). The first rays penetrating the inner core are sharply refracted and emerge in the P-wave shadow zone. The most strongly deviated (ray I) is observed at an epicentral distance of about 110°; deeper rays (J–P) arrive at ever-greater distances up to 180°. There are two branches of the *t*–Δ curve for Δ > 143°, corresponding to the PKP and PKIKP phases, respectively (Fig. 7.28).

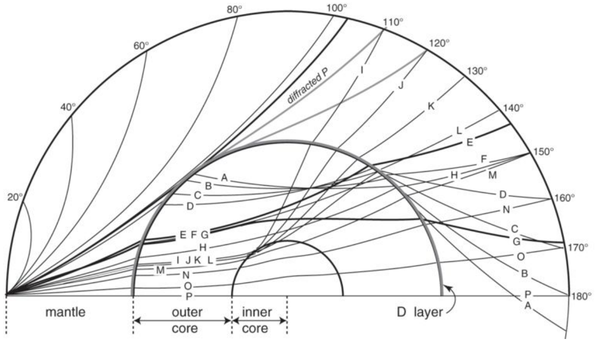


Fig. 7.30 The wave paths of some P, PKP, and PKIKP rays.

The edges of the shadow zone defined by P and PKP phases are not sharp. One reason is the intrusion of PKIKP phases at the 143° edge. Another is the effect of diffraction of P-waves at the 103° edge (Fig. 7.30). The bending of plane waves at an edge into the shadow of an obstruction was described in Section 6.3.2.3, and explained with the aid of Huygens' principle. The diffraction of plane waves is called *Fraunhofer diffraction*. When their source is not at infinity, waves must be handled as spherical waves. Spherical wavefronts that pass an obstacle are also diffracted. This type of behavior is called *Fresnel diffraction*, and it is also explainable with Huygens' principle as the product of interference between the primary wavefront and secondary waves generated at the obstacle. Wave energy penetrates into the shadow of the obstacle, as though the wavefront were bent around the edge. In this way very deep P-waves are diffracted around the core and into the shadow zone. The intensity of the diffracted rays falls off with increasing angular distance from the diffracting edge, in this case the core–mantle boundary. Modern instrumentation enables detection of long-period diffracted P-waves to large epicentral distances (Figs. 7.28, 7.30). The velocity structure above the core–mantle boundary, in particular in the D″-layer (Section 7.4.4.2), has a strong influence on the ray paths, travel times, and waveforms of the diffracted waves.

7.4.2.3 Reverberations within Shallow Layers

When seismic waves that have traveled deep into the mantle approach the surface, they are transmitted through and reflected off several shallow discontinuities. They include, for example, the Moho, the boundary between lower and upper crust, or the top and bottom boundaries of subducted lithosphere near a convergent plate boundary. Multiple reflections and conversions between P-waves and S-waves produce a complicated wave train following the direct P- and S-wave arrivals.

For large epicentral distances, P-waves impinge on a shallow discontinuity with a small angle of incidence measured from the normal (Figs. 7.30, 7.31). Part of the P-wave is transmitted through the discontinuity and continues to travel as a P-wave toward a receiver. Since the upper layer usually has a lower P-wave velocity than the underlying material, the wave will reach the surface nearly vertically. As a consequence, the transmitted P-wave is recorded almost exclusively in the vertical component (Fig. 7.31a). Another part of the original P-wave is converted into an S-wave. Since the transversely polarized S-wave also propagates nearly vertically, it is recorded mostly on the horizontal component (Fig. 7.31b). Once seismic energy has reached the shallow layer, it can reflect many times at its lower and upper boundary, each time producing a new pair of P- and S-waves (Fig. 7.31c).

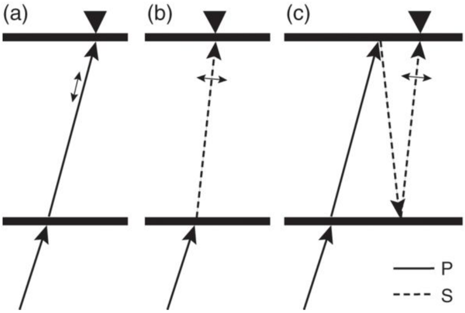


Fig. 7.31 Transmission and conversion of an incoming P-wave at a horizontal discontinuity. The triangle marks the position of a receiver at the surface. Gray arrows indicate the polarization direction of P-waves (solid) and S-waves (dashed) reaching the receiver. (a) A P-wave is transmitted through the discontinuity and continues to propagate as a P-wave toward the receiver. (b) Part of the P-wave energy is converted into an S-wave, primarily recorded on the horizontal component. (c) Waves may reflect multiple times within the layer above the discontinuity, converting into a large variety of P-waves and S-waves.

The horizontal-component S-waves that follow the first-arriving P-wave only exist as a result of conversions at the interface between the lower and upper layers. They carry valuable information about the depth of the interface, but also about its sharpness and the ratio between P-wave and S-wave velocity inside the layer. The use of near-surface P-to-S conversions to study shallow Earth structure is known as *receiver function analysis*.

7.4.3 Inversion of Seismic Data

Models of the radial variations of physical parameters inside the Earth implicitly assume spherical symmetry. They are therefore “average” models of the Earth that do not take into account lateral variations (e.g., of velocity or density) at the same depth. This is a necessary first step in approaching the true distributions as long as lateral variations are relatively small. This appears to be the case. Although geophysically significant, the lateral variations in physical properties mostly remain within a few percent of the average value at any depth below the crust.

In general, there are two approaches to determine physical properties, such as seismic velocities, inside the Earth. They are referred to as *direct inversion* and *indirect inversion*. Both methods employ the same set of observations, for example, the travel times of different seismic phases to known epicentral distances. Direct inversion means that the observations can be inserted directly into an equation which returns a model of the Earth’s properties. Though direct inversion is elegant and efficient, an equation that performs the inversion can only be found in a few exceptional cases. Furthermore, observational uncertainties, such as errors in picking the travel times of seismic waves, are difficult to incorporate in direct inversion methods.

Indirect inversion is based on solutions of the *forward problem*, that is, the prediction of observations using a model of the Earth. Examples are the computations of seismic travel times for layered media in Section 6.3. Starting from an initial guess, indirect inversion solves the forward problem repeatedly until it finds a model of the Earth that matches the observations.

7.4.3.1 Direct Inversion of Travel Time versus Distance Curves

In 1907 the German geophysicist Emil Wiechert, building upon an evaluation of the Benndorf relationship (Eq. 6.93) by the mathematician Gustav Herglotz, developed a direct inversion method for computing the internal distributions of seismic velocities from travel time observations made at the Earth’s surface. The data consist of the t–Δ curves for important seismic phases (Fig. 7.28). The clues to deciphering the velocity distributions were the Benndorf relationship for the ray parameter p (Eq. 6.93) and the recognition that the value of p can be obtained from the slope of the travel time curve at the epicentral distance where the ray returns to the surface.

Consider two rays that leave an earthquake focus at infinitesimally different angles, reaching the surface at points P and P′ at epicentral distances Δ and $\Delta+\mathrm{d}\Delta$, respectively (Fig. 7.32). The distance PP′ is $R\,\mathrm{d}\Delta$ (where R is the Earth’s radius) and the difference in arc-distances along the adjacent rays is equal to $\alpha\,\mathrm{d}t$, where α is the P-wave velocity in the surface layer, and $\mathrm{d}t$ is the difference in travel times of the two rays. (The same arguments are valid for S-waves when α is replaced by the S-wave velocity β.) In the small triangle PP ′Q the angle QPP′ is equal to i, the angle of emergence (incidence). Therefore,

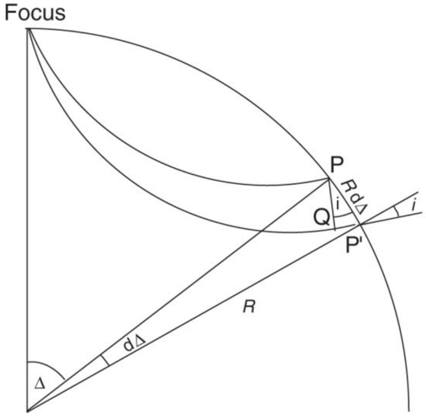


Fig. 7.32 Paths of two rays that leave an earthquake focus at infinitesimally different angles, reaching the surface at points P and P′ at epicentral distances Δ and $\Delta+\mathrm{d}\Delta$, respectively.

$$\sin i = \frac{\alpha \mathrm{d}t}{R \mathrm{d}\Delta}, \tag{7.18}$$

$$\frac{R \sin i}{\alpha} = p = \frac{\mathrm{d}t}{\mathrm{d}\Delta}. \tag{7.19}$$

This means that the value of *p* for the ray emerging at epicentral distance Δ can be obtained by calculating the slope (d*t*/dΔ) of the travel time curve for that distance. This is an important step in finding the velocity α_0 at the deepest point of the ray, at radius r_0, because $\alpha_0 = r_0/p$. However, before we can find the velocity we need to know the value of r_0. The continuation of the analysis – known as Herglotz–Wiechert inversion – is an intricate mathematical procedure, beyond the scope of this text, which fortunately results in a fairly simple formula:

$$\ln\left(\frac{R}{r_0}\right) = \frac{1}{\pi} \int_0^{\Delta_1} \cosh^{-1}\left(\frac{p(\Delta)}{p(\Delta_1)}\right) \mathrm{d}\Delta, \tag{7.20}$$

where $p(\Delta_1)$ is the slope of the *t*–Δ curve at Δ_1, the epicentral distance of emergence, and $p(\Delta)$ is the slope at any intermediate epicentral distance Δ. Equation (7.20) can be used to integrate numerically along the ray to give the value of r_0 for the ray, and $\alpha_0 = r_0/p$.

The Herglotz–Wiechert method is one of the few cases in seismology where a direct inversion equation can be found. It assumes that the seismic velocity distribution in the Earth is indeed nearly spherically symmetric. Furthermore, the derivation of Eq. (7.20) requires that seismic velocity increases monotonically with depth, meaning that no low-velocity zones are present. While these constraints limit the practical applicability of the Herglotz–Wiechert equation, it still serves as a useful illustration of the direct inversion approach.

7.4.3.2 Indirect Inversion of Travel Times and Seismic Tomography

The indirect inversion approach circumvents the limitations of direct inversion and is also applicable when a direct inversion equation cannot be found, for example in the presence of low-velocity layers or three-dimensional heterogeneities. This flexibility usually comes at the price of higher computational costs. Indirect inversion starts with a presupposed dependence of seismic velocity with depth. The travel times of P- and S-waves to a set of receivers are then calculated and compared with the observations. Subsequently, the velocity model is adjusted to account for the differences.

Seismic tomography is a special variant of indirect inversion. A free translation of the term tomography is “representation in cross-section.” Neighboring two-dimensional cross-sections can be combined to give a three-dimensional model. Computer-aided tomography (CAT) in medical diagnosis is a well-known form of tomography for the non-invasive examination of internal organs. X-rays or ultrasonic rays travel through tissue with variable physical properties. Computer-aided tomography constructs an image of these properties for use in medical diagnosis. Seismic tomography uses similar principles, with the difference that the travel times and the attenuation of seismic waves are observed.

The travel time of a seismic wave from a source to a receiver is determined by the velocity distribution along its path. For an idealized, spherically symmetric Earth model the radial distribution of velocity is known. The velocities in the model are mean values, which average out lateral fluctuations. If such a velocity model were used to compute travel times of different phases to any epicentral distance, a set of curves indistinguishable from Fig. 7.28 would result. In reality, the observed travel times usually show small deviations from the calculated times, known as *travel time residuals*. They can have several causes, including a mislocation of the source, ellipticity of the Earth, topography, and lateral variations in P-wave and S-wave velocities somewhere along the path.

If, at a certain depth, a seismic ray passes through a region in which the velocity is slightly faster than the radial average, the wave will arrive slightly sooner than expected at the receiver; if the anomalous velocity is slower than average, the wave will arrive late.

Box 7.2 Calculation of Travel Time Residuals

Consider the passage of seismic rays in six directions through a square region containing four equal areas, each a square with edge 5 km, in which the P-wave velocities (α) are 4.9 km s^{-1}, 5.0 km s^{-1}, 5.1 km s^{-1}, and 5.2 km s^{-1} (Fig. B7.2.1). Let the expected average velocity (α_0) throughout the region be 5.0 km s^{-1}. The velocity difference ($\Delta\alpha$) in each area is found by subtracting the reference value, and the velocity anomaly ($\Delta\alpha/\alpha_0$) is obtained by expressing the difference as a percentage of the expected value. This gives zones that are 2% and 4% fast, a zone that is 2% slow, and a zone with no anomaly.

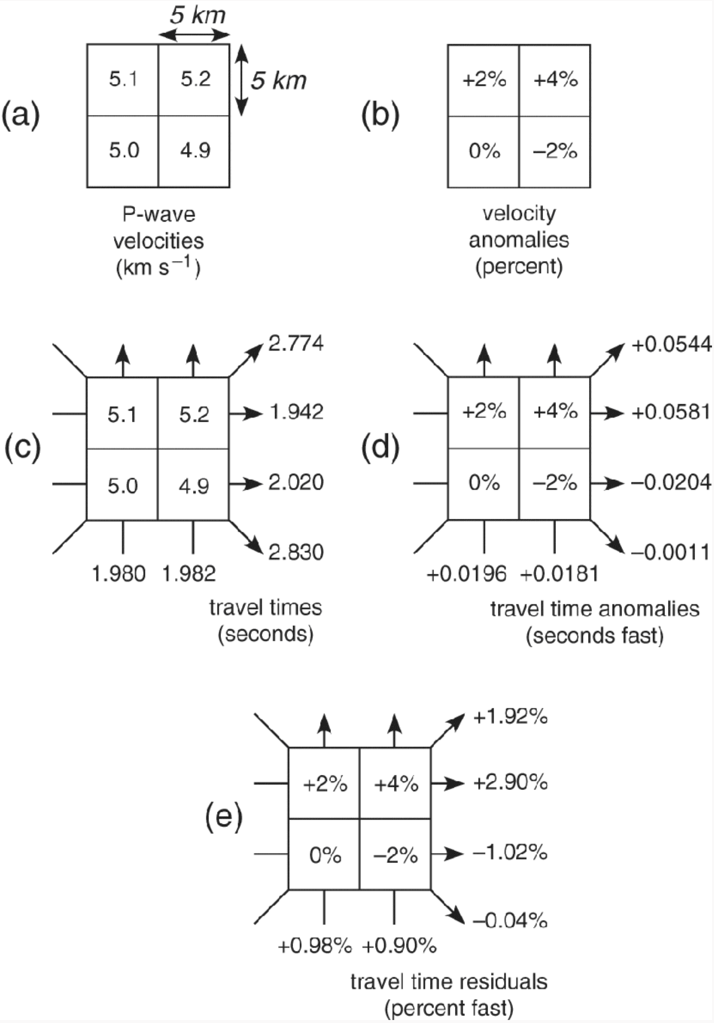

5 km
5 km
(a)
5.1
5.2
5.0
4.9
P-wave velocities (km s−1)
(b)
+2%
+4%
0%
−2%
velocity anomalies (percent)
(c)
5.1
5.2
5.0
4.9
2.774
1.942
2.020
2.830
1.980
1.982
travel times (seconds)
(d)
+2%
+4%
0%
−2%
+0.0544
+0.0581
−0.0204
−0.0011
+0.0196
+0.0181
travel time anomalies (seconds fast)
(e)
+2%
+4%
0%
−2%
+1.92%
+2.90%
−1.02%
−0.04%
+0.98%
+0.90%
travel time residuals (percent fast)

Fig. B7.2.1 Computation of relative travel-time residuals (in percent) for a simple four-block structure.

Source: E. Kissling, Seismische Tomographie: Erdbebenwellen durchleuchten unseren Planeten, *Vierteljahresschrift Naturforsch. Ges. Zürich* 138, 1–20, 1993.

Suppose that six seismic rays traverse the square region as in Fig. B7.2.1. If the velocity were 5.0 km s^{-1} in each area, the expected travel time (t_0) would be 2.0 s for each of the horizontal and vertical rays and $2\sqrt{2} = 2.828$ s for the longer diagonal rays. However, the real velocity is different in each area. As a result, some of the observed travel times are shorter and some are longer than the expected value. A travel time anomaly (Δt) is computed by subtracting the observed travel-times (t) from the expected value (Fig. B7.2.1d). The *travel time residual* ($\Delta t/t_0$) is obtained by expressing the anomaly as a percentage of the expected travel time (Fig. B7.2.1e). Note that the residuals are not simple averages of the velocity anomalies along each path.

The time (t) taken for a seismic ray to travel along the path from a source to a recording station thus differs from the value (t_0) predicted by the model. The difference ($\Delta t = t_0 - t$) is a *travel time anomaly*. A *travel time residual* ($\Delta t/t_0$) for the path is obtained by expressing the anomaly as a percentage of the expected travel-time (Box 7.2). If a given path is subdivided into segments with different velocities, the travel time is equivalent to the sum of the travel times through the individual segments. This is like an equation with one known value (Δt) and several unknown terms (the velocity anomalies in each segment). For a given source several seismic stations may record rays that traverse the region of interest. Each station may also have recorded rays that crossed the region from different sources. A large set of travel time residuals results, equivalent to a large number of equations. A mathematical procedure called matrix inversion is used to solve the set of equations and obtain the *velocity anomalies*. The analysis involves intensive data processing, and is beyond the scope of this text. However, Box 7.3 illustrates how a distribution of velocity anomalies may be deduced by *back projection* for a simple pattern of travel time residuals (Box 7.2). In this method successive adjustments are made to an initial velocity distribution to account for the observed travel time residuals.

Box 7.3 Calculation of Velocity Anomalies

To illustrate how a velocity structure can be deduced from observed travel time residuals, we take as a starting point the set of travel time residuals ($\Delta t/t_0$) computed for the velocity distribution in Box 7.2. The expected P-wave travel time along a seismic path of length L is $t_0 = L/\alpha_0$, where α_0 is the reference velocity. If the true velocity is $\alpha = \alpha_0 + \Delta\alpha$, the observed travel time is

$$\begin{aligned} t = \frac{L}{\alpha_0 + \Delta\alpha} &= \frac{L}{\alpha_0}\left(1 + \frac{\Delta\alpha}{\alpha_0}\right)^{-1} \\ &= t_0\left(1 - \frac{\Delta\alpha}{\alpha_0} + \left(\frac{\Delta\alpha}{\alpha_0}\right)^2 + \cdots\right), \end{aligned}$$

(1)

$$\left(\frac{\Delta t}{t_0}\right) = \left(\frac{\Delta\alpha}{\alpha_0}\right)\left(1 - \frac{\Delta\alpha}{\alpha_0} + \cdots\right).$$

(2)

Next, we assume that the percentage P-wave velocity anomalies ($\Delta\alpha/\alpha_0$) are the same as the percentage travel time residuals. As shown by Eq. (2), this is not strictly true; the discrepancy is of the order of ($\Delta\alpha/\alpha_0$). However, when the velocity anomalies are very small, assuming that the velocity anomaly is equal to the travel time residual is a reasonable first approximation.

Consider a horizontal ray that traverses the upper two “fast” blocks in Fig. B7.3.1. To account for the (early) travel time anomaly of +2.9%, let each upper block be allocated a (fast) velocity anomaly of +2.9%. Similarly, let each of the bottom two blocks be allocated velocity anomalies of –1.02%. This simple velocity distribution (Fig. B7.3.1a) satisfies the travel time anomalies for the horizontal rays. However, in the vertical direction it gives travel time anomalies of +0.94% (the mean of +2.90% and –1.02%) for each ray (Fig. B7.3.1b). This does not agree with the observed anomalies for the two vertical rays (+0.98% and +0.90%, respectively); one vertical anomaly is 0.04% too large, the other is 0.04% too small.

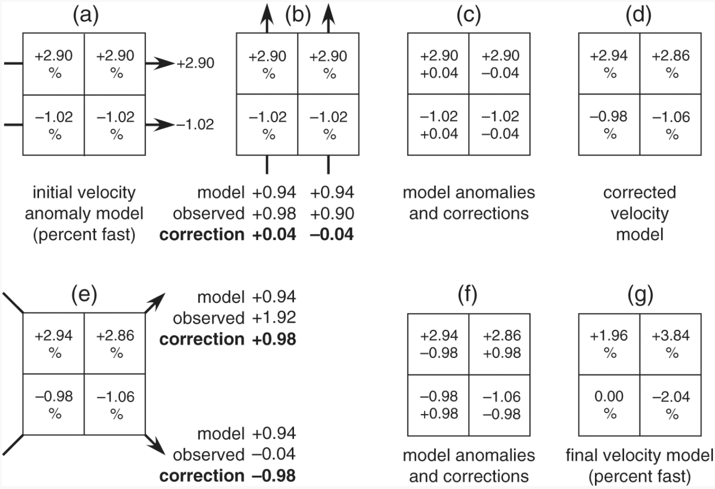

(a)
+2.90 %
+2.90 %
+2.90
–1.02 %
–1.02 %
–1.02
initial velocity anomaly model (percent fast)
(b)
+2.90 %
+2.90 %
–1.02 %
–1.02 %
model +0.94 +0.94
observed +0.98 +0.90
correction +0.04 –0.04
(c)
+2.90 +0.04
+2.90 –0.04
–1.02 +0.04
–1.02 –0.04
model anomalies and corrections
(d)
+2.94 %
+2.86 %
–0.98 %
–1.06 %
corrected velocity model
(e)
+2.94 %
+2.86 %
–0.98 %
–1.06 %
model +0.94
observed +1.92
correction +0.98
model +0.94
observed –0.04
correction –0.98
(f)
+2.94 –0.98
+2.86 +0.98
–0.98 +0.98
–1.06 –0.98
model anomalies and corrections
(g)
+1.96 %
+3.84 %
0.00 %
–2.04 %
final velocity model (percent fast)

Fig. B7.3.1 Backward projection of relative travel time residuals to obtain velocity anomalies.

Source: E. Kissling, Seismische Tomographie: Erdbebenwellen durchleuchten unseren Planeten, *Vierteljahresschrift Naturforsch. Ges. Zürich* 138, 1–20, 1993.

The velocities in the blocks are adjusted accordingly by making a correction of +0.04% to the left-hand blocks and –0.04% to the right-hand blocks (Fig. B7.3.1c). This gives a new distribution of velocity anomalies which satisfies the horizontal and vertical rays (Fig. B7.3.1d).

The model now gives travel time anomalies along each of the diagonal rays of +0.94%, compared to observed anomalies of +1.92% and –0.04%, respectively (Fig. B7.3.1e). Further corrections of +0.98% are now made to the upper right and lower left blocks, and –0.98% to the lower right and upper left blocks (Fig. B7.3.1f). The resulting distribution of velocity anomalies (Fig. B7.3.1g) satisfies all six rays through the anomalous region, and is close to the original distribution of velocity anomalies (Fig. B7.2.1b).

Tomographic imaging of the Earth’s three-dimensional velocity distribution is able to resolve small differences in seismic velocities. In some studies the signals are of local origin, generated by earthquakes or explosions within or near the volume of interest. Other studies are based on teleseismic signals, which originated in earthquakes at more than 20° epicentral distance. The analysis requires precise location of the earthquake source, corrections for local effects (such as crustal structure close to each measurement station), and the calculation of the seismic rays between source and seismometer. Seismic tomography can be based on either body waves or surface waves. The velocity structures of both P-waves and S-waves have been tomographically imaged for global models of the entire mantle as well as for regional studies (see Section 7.4.6).

7.4.3.3 Seismic Waveform Tomography

The three-dimensional structure of the Earth influences not only the travel time of a seismic wave but also its amplitude at each instant of time. Methods known as *waveform tomography* use this additional information by matching complete simulated and observed seismograms. Being an indirect inversion method, waveform tomography is based on repeated forward simulations of seismic wave propagation, mostly using the numerical methods described in Section 6.3.7. Starting with an initial guess of Earth structure, a first set of simulated seismograms is computed. The simulated seismograms and the actually observed seismograms usually do not agree well. Some waves may arrive too early, others too late, or the amplitudes of some waves may not match. Following a suitable modification of the initial guess, a new set of seismograms is simulated and compared to the observations. This procedure is then iterated until an Earth model that explains the observed seismograms is found. To limit the number of necessary iterations, numerous methods have been developed for the efficient improvement of Earth models. These are, however, largely beyond the scope of this book.

An example of waveform tomography in the eastern Mediterranean is shown in Fig. 7.33. The three-dimensional distribution of S-wave velocity produces simulated seismograms that match complete seismic recordings. This includes the travel times and the amplitudes of early-arriving body waves, but also of the later-arriving surface waves. The ability to include information from all seismic wave types is one of the main advantages of waveform tomography, which is still an emerging technology that requires modern supercomputers.

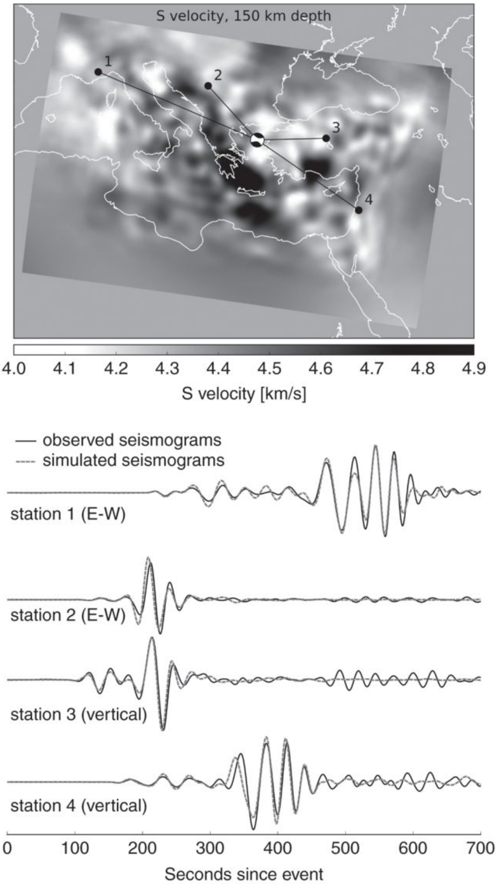


Fig. 7.33 Illustration of waveform tomography in the eastern Mediterranean. The S-wave velocity structure, only shown at 150 km depth, has been determined by indirect inversion of several thousand seismogram recordings. A comparison of observed and simulated seismograms at four stations is shown for one earthquake, marked by the beachball. The travel times and the amplitudes of body waves (arriving early, with lower amplitudes) and also of surface waves (arriving later, with higher amplitudes) are well-matched. (Courtesy of Nienke Blom.)

7.4.4 Radial Variations of Seismic Velocities in the Earth

While the Earth's crust is too heterogeneous to be represented in a meaningful way by some radial average, this changes below the Moho, where seismic velocities typically vary by no more than a few percent. Since high-quality seismic data with global coverage became available in the 1970s, several spherically symmetric Earth models have been constructed using indirect inversion. In 1981 Adam M. Dziewonski and Don L. Anderson constructed a *Preliminary Reference Earth Model* (PREM) in which the distributions of body-wave velocities in the Earth are represented by polynomials. In the 1990s, revised velocity models (*iasp91* and *ak135*) were proposed by Brian L. N. Kennett, E. Robert Engdahl, and Raymond Buland. The variations of P- and S-wave velocities with depth in the Earth according to the *iasp91* model are shown in Fig. 7.34. Though spherical Earth models constructed during the past few decades differ in small details, they show a very consistent picture of average Earth structure, discussed in the following subsections.

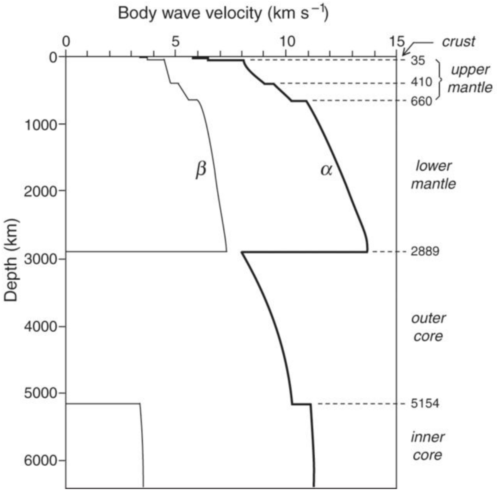


Fig. 7.34 The variations with depth of longitudinal- and shear-wave velocities, α and β, respectively, in the Earth's interior, according to the Earth model *iasp91* .

Data source: B.L.N. Kennett and E.R. Engdahl, Traveltimes for global earthquake location and phase identification, *Geophys. J. Int.* 105, 429–465, 1991.

7.4.4.1 The Upper Mantle

The upper mantle in bounded by two major abrupt changes in seismic velocities, the Moho and the 660 km discontinuity, where the mantle mineral spinel transforms to perovskite. The composition of the upper mantle is mostly peridotitic, with olivine [$(Mg,Fe)_2SiO_4$] as the dominant mineral. With increasing depth the lithostatic pressure increases and eventually causes high-pressure transformations of the silicate minerals. This is reflected in the seismic properties. Travel time–Δ curves of body waves show a distinct change in slope at epicentral distances of about 20°. This is attributed to another discontinuity in mantle velocities at a depth of around 410 km (Fig. 7.34). The 410 km discontinuity is interpreted as a mineral phase change from an olivine-type lattice to a more closely packed spinel-type lattice. In the *transition zone* between the 410 km and 660 km discontinuities there is a further change in structure from β-spinel to γ-spinel, but this is not accompanied by large changes in physical properties.

In addition to the globally observed discontinuities at 410 km and 660 km depth, the upper mantle also contains other features that may not always be present. This includes a pronounced low-velocity zone (LVZ) and a discontinuity around 220 km depth, called the *Lehmann discontinuity*. Whether these features should be part of a model of spherically symmetric Earth structure is a matter of personal choice.

While the nature of the Lehmann discontinuity is still debated, the LVZ between roughly 100 and 200 km depth is commonly interpreted as a seismic expression of the asthenosphere, which plays an important role in plate tectonics. The decrease in seismic velocities is attributed to reduced rigidity in this layer. Over geological time intervals the mantle reacts like a viscous medium, with a viscosity that depends on temperature and composition. From the point of view of plate tectonics the asthenosphere is a viscous layer that decouples the lithosphere from the deeper mantle; by allowing slow convection, it permits or promotes the relative motions of the global plates.

7.4.4.2 The Lower Mantle

The lower mantle is bounded by the 660 km discontinuity and the core–mantle boundary at 2889 km depth. It consists of oxides of iron and magnesium as well as iron–magnesium silicates with a perovskite structure. Just above the core–mantle boundary a layer, approximately 150–200 km thick, has been identified in which velocity gradients are very small and may even be negative. Although part of the lower mantle, it evidently serves as a boundary layer between the mantle and the core. It is commonly labeled D″, in reminiscence of an Earth model proposed by Keith E. Bullen, who labeled layers alphabetically. The detailed structure and the nature of D″ are the subject of intensive research. It likely plays an important role in geodynamic and geothermal behavior. On the one hand, D″ may serve as the source of material for mantle plumes that give rise to hotspots, which are important in plate tectonics. On the other hand, the thermal properties of D″ could influence the outward transport of heat from the Earth's core; in turn, this could affect the intricate processes that generate the Earth's magnetic field.

7.4.4.3 The Core

Early in Earth's history, dense metallic elements are thought to have settled toward the Earth's center, forming the core, while lighter silicates ascended and solidified to form the mantle. Studies of the compositions of meteorites and of the behavior of metals at high pressure and temperature give a plausible picture of the composition and formation of the core. It consists mainly of iron, with perhaps up to 10% nickel. The observed pressure–density relationships suggest that some less-dense non-metallic elements (Si, S, O) may be present in the outer core.

The core has a radius of 3480 km and consists of a solid inner core surrounded by a liquid outer core that is 1220 km thick. Improved seismic instrumentation has yielded a large quantity of high-quality data for reflections from the boundaries of the outer core (PcP, ScS) and the inner core (PKiKP) which have helped to clarify the nature of these boundaries.

The core–mantle boundary is also called the *Gutenberg discontinuity*. It is characterized by large changes in seismic velocities and is the most sharply defined seismic discontinuity. Seismic data show that the boundary is not smooth, but has a topography of hills and valleys. Anomalies in the travel times of PKKP phases – which are reflected once internally at the core–mantle boundary – have been attributed to scattering by topographic features with a relief of a few hundred meters. However, depending on conditions in the D″ layer, topographic features may be up to 10 km high.

The PKiKP phase contains high frequencies, suggesting that the inner core boundary is sharp, probably no more than 5 km thick. The outer core is fluid, with a viscosity similar to that of water. For seismic waves, the outer core behaves like a homogeneous medium, which is expected in a well-mixed fluid.

The inner core transmits P-waves (PKIKP phase) and S-waves (PKJKP phase), although the latter are much more difficult to observe. Observations of PKJKP and of the Earth’s normal modes show that the inner core is solid. It forms through the solidification (freezing) of fluid outer core material, leaving behind its lighter elements. These constitute a less-dense, therefore gravitationally buoyant, fluid, which rises through the denser overlying liquid. This compositional type of buoyancy contributes to convection in the outer core, and therefore to the dynamics of the core and generation of the Earth’s magnetic field.

7.4.5 Radial Variations of Density, Gravity, and Pressure

While the spherically symmetric distribution of seismic velocities is well constrained by travel time observations at the surface, the density structure of the Earth is less well known. This is because seismic waves are only weakly sensitive to variations in density, and because gravity observations generally do not provide unambiguous constraints on density structure.

However, with some simplifications, the variations of density with depth can be estimated from the spherically symmetric velocity variations. For this, the Earth is assumed to be composed of concentric homogeneous shells or layers (e.g., inner core, outer core, mantle, etc.). Possible effects of chemical and phase changes within a shell are ignored, and pressure and density are assumed to increase purely hydrostatically. If the distributions of seismic body-wave velocities α and β are known, an important seismic parameter Φ can be defined:

$$\Phi = \alpha^2 - \frac{4}{3}\beta^2. \tag{7.21}$$

By comparing Eq. (7.21) and Eq. (6.16), we see that Φ is equal to K/ρ, where K is the bulk modulus and ρ the density.

7.4.5.1 Density Inside the Earth

Consider a vertical prism between depths z and dz (Fig. 7.35). The hydrostatic pressure increases from p at depth z to (p + dp) at depth (z + dz) because of the extra pressure due to the extra material in the small prism of height dz. The pressure increase dp is equal to the weight w of the prism divided by the area A of the base of the prism, over which the weight is distributed:

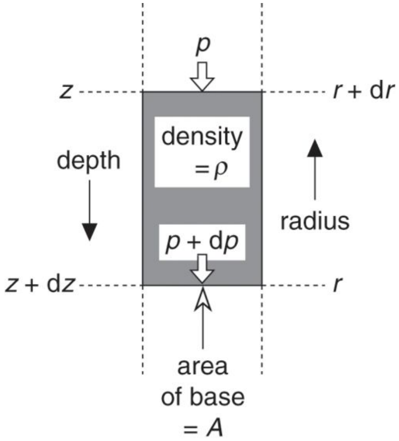


Fig. 7.35 Computation of hydrostatic pressure in the Earth, assuming that a change in pressure d*p* with depth increase d*z* is due only to the increase in weight of overlying material.

$$\mathrm{d}p = \frac{w}{A} = \frac{(\text{volume} \times \rho)}{A} = \frac{(A\mathrm{d}z\rho)}{A} = \rho g \mathrm{d}z = -\rho g \mathrm{d}r. \tag{7.22}$$

From the definition of bulk modulus, *K* (Eq. 5.17) we can write

$$K = -V\frac{\mathrm{d}p}{\mathrm{d}V} = \rho\frac{\mathrm{d}p}{\mathrm{d}\rho}. \tag{7.23}$$

Combining Eqs. (7.21), (7.22), and (7.23) gives

$$\frac{-\mathrm{d}\rho}{\rho g \mathrm{d}r} = \frac{\rho}{K} = \frac{1}{\Phi}, \tag{7.24}$$

$$\frac{\mathrm{d}\rho}{\mathrm{d}r} = -\frac{\rho(r)g(r)}{\Phi(r)}. \tag{7.25}$$

Equation (7.25) is known as the Adams–Williamson equation. It was first applied to the estimation of density in the Earth in 1923 by Leason H. Adams and Erskin D. Williamson. It yields the density gradient at radius r, when the quantities on the right-hand side are known. The seismic parameter Φ is known accurately, but the density ρ is unknown; it is in fact the goal of the calculation. The value of gravity g used in the equation must be computed separately for radius r. It is due only to the mass contained within the sphere of radius r, because external (homogeneous) shells of the Earth do not contribute to gravitation inside them. This mass is the total mass M of the Earth minus the cumulative mass of all spherical shells external to r.

The procedure requires that a starting value for ρ be assumed at a known depth. Analyses of crustal and upper mantle structure in isostatically balanced areas give estimates of upper mantle density around 3300 kg m^{-3}. Using this as a starting value at an initial radius r_1, the density gradient in the uppermost mantle can be calculated from Eq. (7.25). Linear extrapolation of this gradient to a chosen greater depth gives the density ρ at radius r_2; the corresponding new value of g can be calculated by subtracting the mass of the shell between the two depths; together with the value of $\Phi(r_2)$ the density gradient can be computed at r_2. This iterative type of calculation gives the variation of density with depth (or radius). Fine steps in the extrapolations give a smooth density distribution (Fig. 7.36). There are two important boundary conditions on the computed density distribution. Integrated over the Earth's radius it must give the correct total mass ($M = 5.974 \times 10^{24}$ kg). It must also fulfill the relationship $C = 0.3308\, MR^2$ between Earth's moment of inertia (C), mass (M), and radius (R) (see Section 1.1.1).

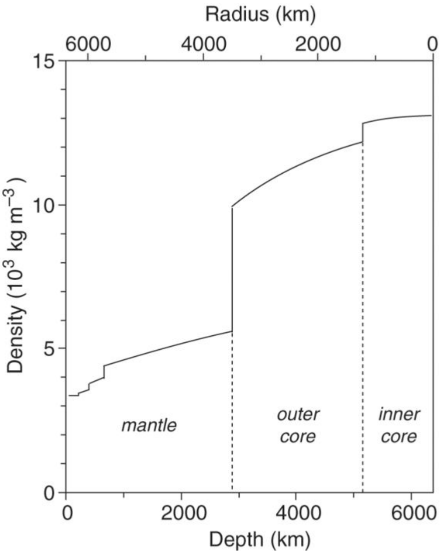


Fig. 7.36 Radial distribution of density within the Earth according to Earth model PREM (data source: Dziewonski, 1989).

Density changes abruptly at the major seismic discontinuities (Fig. 7.36), showing that it is affected principally by changes in composition. If it could be measured at normal sea-level pressure and temperature, density would be found to be around 4200 kg m^{-3} in the mantle, 7600 kg m^{-3} in the outer core, and 8000 kg m^{-3} in the inner core. The smooth increase of density between the major compositional discontinuities is the result of the increases in pressure and temperature with depth.

7.4.5.2 Gravity and Pressure Inside the Earth

The radial variation of gravity can be computed from the density distribution. As stated above, the value of $g(r)$ is due only to the mass $m(r)$ contained within the sphere of radius r. Let the density at radius x ($\leq r$) be $\rho(x)$; the gravity at radius r is then given by

$$g(r) = -G\frac{m(r)}{r^2} = -\frac{G}{r^2}\int_0^r 4\pi x^2 \rho(x)\mathrm{d}x. \qquad (7.26)$$

A remarkable feature of the internal gravity (Fig. 7.37) is that it maintains a value close to 10 m s^{-2} throughout the mantle, merely varying from 9.8 m s^{-2} at the surface to 10.8 m s^{-2} at the core–mantle boundary. It then decreases almost linearly to zero at the Earth's center.

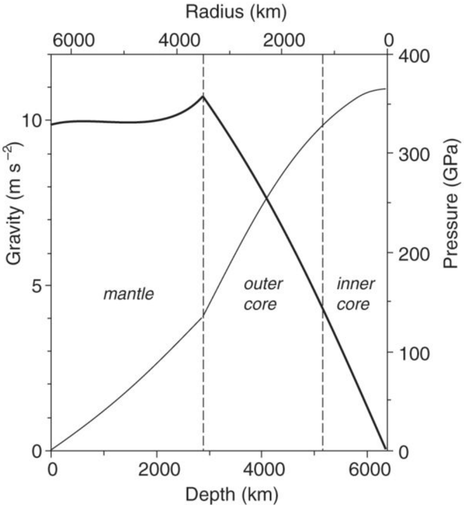


Fig. 7.37 Radial variations of internal gravity (*thick curve*) and pressure (*thin curve*) according to Earth model PREM (data source: Dziewonski, 1989).

Hydrostatic pressure is due to the force (N) per unit area (m^2) exerted by overlying material. The SI unit of pressure is the pascal (1 Pa = 1 N m^{-2}). In practice this is a small unit. The high pressures in the Earth are commonly quoted in units of gigapascal (1 GPa = 10^9 Pa), or alternatively in kilobars or megabars (1 bar = 10^5 Pa; 1 kbar = 10^8 Pa; 1 Mbar = 10^{11} Pa = 100 GPa).

Within the Earth the hydrostatic pressure $p(r)$ at radius r is due to the weight of the overlying Earth layers between r and the Earth's surface. It can be computed by integrating Eq. (7.22) using the derived distributions of density and gravity. This gives

$$p(r) = \int_{r}^{R} \rho(r) g(r) \mathrm{d}r.$$

(7.27)

Pressure increases continuously with increasing depth in the Earth (Fig. 7.37). The rate of increase (pressure gradient) changes at the depths of the major seismic discontinuities. The pressure reaches a value close to 380 GPa (7.8 Mbar) at the center of the Earth, which is about four million times atmospheric pressure at sea level.

7.4.6 Three-Dimensional Earth Structure and Its Relation to Geodynamic Processes

The construction of models that describe the average radial structure of the Earth with high accuracy enables the discovery of three-dimensional heterogeneities in seismic velocities and density. Such heterogeneities are the results of variations in temperature and composition. They provide information on the current deformation and evolution of the Earth's internal structure.

7.4.6.1 The Crust

The most apparent departure from a spherically symmetric Earth model is the Earth's crust. Detailed information on crustal seismic structure comes from seismic refraction profiles and deep crustal reflection sounding. It reveals that ancient continental shields have different vertical velocity profiles than younger continental or oceanic domains. In view of this variability, any generalized petrological model of crustal structure is bound to be an oversimplification. However, with this reservation, it is still possible to summarize some general features of crustal structure, and the corresponding petrological layering.

A generalized model of the structure of oceanic crust is shown in Fig. 7.38. Oceanic crust is less than 10 km thick. Under a mean water depth of about 4.5 km, the top part of the oceanic crust consists of a layer of sediments (Layer 1) that increases in thickness away from the oceanic ridges. Beneath the sediments the igneous oceanic basement consists of two layers. A thin (~0.5 km) upper layer of superposed basaltic lava lies on top of a complex of basaltic intrusions, the sheeted dike complex; together these form oceanic Layer 2. Below the basaltic layer the oceanic crust consists of gabbroic rocks (Layer 3). The direct seismic wave in Layer 2 is labeled P_2, the head wave at the basalt–gabbro interface is referred to as P_3, and the Moho head wave is labeled P_n.

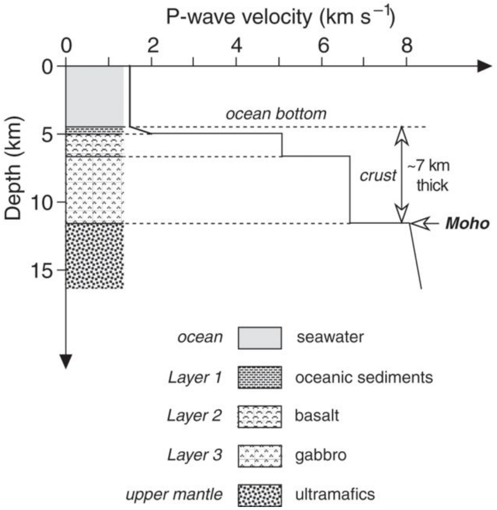

P-wave velocity (km s^{-1})
0
2
4
6
8
Depth (km)
0
5
10
15
ocean bottom
crust
~7 km thick
Moho
ocean
seawater
Layer 1
oceanic sediments
Layer 2
basalt
Layer 3
gabbro
upper mantle
ultramafics

Fig. 7.38 Generalized petrological model and P-wave velocity–depth profile for oceanic crust.

A refraction seismic record from the European continental crust is illustrated in Fig. 7.39. The vertical axis in this figure shows a reduced travel time, obtained by dividing the shot-point to receiver distance by a representative crustal velocity ($6\ \mathrm{km\ s^{-1}}$ in this case) and subtracting this from the observed travel time. This method of displaying the data prevents the plot from becoming unwieldy at large distances. The crustal direct wave P_g is represented as a nearly horizontal arrival, PmP is a P-wave reflected from the Moho, and P_n is the upper mantle head wave along the Moho. Additional intracrustal reflections are labeled P2P, P3P, etc.

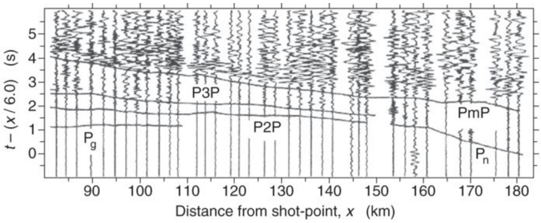


Fig. 7.39 A southwest–northeast refraction seismic profile parallel to the strike of the Swiss Alps, showing reduced travel times as a

function of distance from the shot-point.

Source: H. Maurer and J. Ansorge, Crustal structure beneath the northern margin of the Swiss Alps, *Tectonophysics*, 207, 165–181, 1992.

The vertical structure of continental crust is more complicated than that of oceanic crust, and the structure under ancient shield areas differs from that under younger basins. It is more difficult to generalize a representative model (Fig. 7.40). The most striking difference is that the continental crust is much thicker than oceanic crust. Under stable continental areas the crust is on average 35–40 km thick and under young mountain ranges it is often 50–70 km thick. The continental Moho is not always a sharp boundary. In some places the transition from crust to mantle may be gradual, with a layered structure. Originally, the Conrad seismic discontinuity was believed to separate an upper crustal (granitic) layer from a lower crustal (basaltic) layer. However, a simple Conrad discontinuity is not found in all places. Crustal velocity studies have defined two anomalous zones that often disrupt the otherwise progressive increase of velocity with depth. A low-velocity layer within the middle crust is thought to be due to intruded granitic laccoliths; it is called the sialic low-velocity layer. It is underlain by a middle crustal layer composed of migmatites. Below this layer the velocity rises sharply, forming a “tooth” in the velocity profile. This tooth and the layer beneath it often make up a thinly layered lower crust. Refractions and reflections at the top of the tooth are thought to explain the Conrad discontinuity.

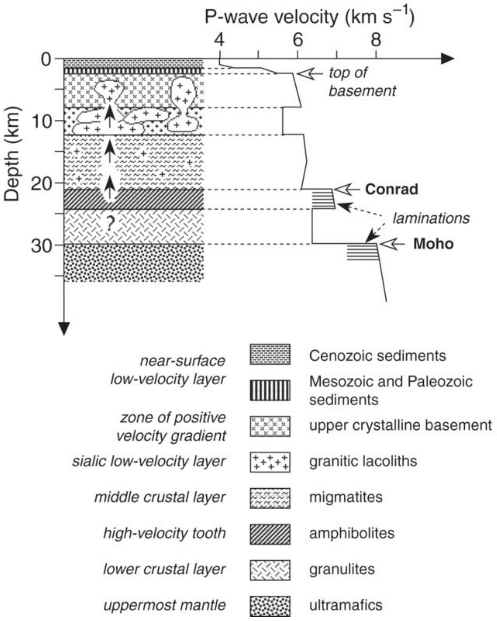

P-wave velocity (km s⁻¹)
4
6
8
0
10
20
30
Depth (km)
?
top of basement
Conrad
laminations
Moho
near-surface low-velocity layer
zone of positive velocity gradient
sialic low-velocity layer
middle crustal layer
high-velocity tooth
lower crustal layer
uppermost mantle
Cenozoic sediments
Mesozoic and Paleozoic sediments
upper crystalline basement
granitic lacoliths
migmatites
amphibolites
granulites
ultramafics

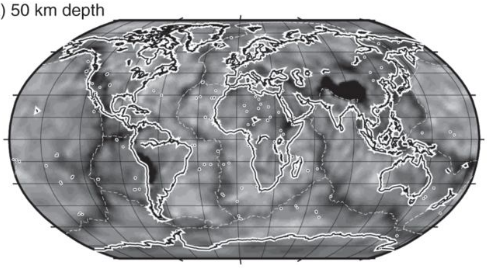
Fig. 7.40 Generalized petrological model and P-wave velocity–depth profile for continental crust.

Source: S. Mueller, A new model of the continental crust, in: *Geophysical Monograph 20: The Earth's Crust*, J.G. Heacock, ed., pp. 289–317, Washington, D.C: American Geophysical Union, 1977.

7.4.6.2 The Upper Mantle

The three-dimensional seismic structure of the Earth's upper mantle can be investigated using the travel times of body waves and the dispersion of surface waves. While body waves may travel deep into the Earth, surface waves are generally confined to the shallower layers. As a rule of thumb, a surface wave with a period of T seconds is sensitive to Earth structure between the surface and a depth of $2T$ kilometers. For example, a 30 s surface wave can "see" structure to a depth of around 60 km.

Seismic velocity heterogeneities in the upper mantle show a strong imprint of plate tectonic processes (Figs. 7.41 and 7.42). Mid-ocean ridges, for example, coincide with a narrow band where S-wave velocity is reduced by around 5% down to 120 km depth or more (Fig. 7.42c,d). Reasons for such velocity variations include changes in temperature or in composition, with the effect of temperature being dominant in the upper mantle. Mineral physics experiments indicate that a relative change of S-wave velocity caused by a change in temperature, $(\Delta\beta/\beta)/\Delta T$, is approximately –2% per 100°C at 100 km depth. This suggests that material beneath mid-ocean ridges is around 250°C hotter than the surrounding upper mantle.


(a) 50 km depth

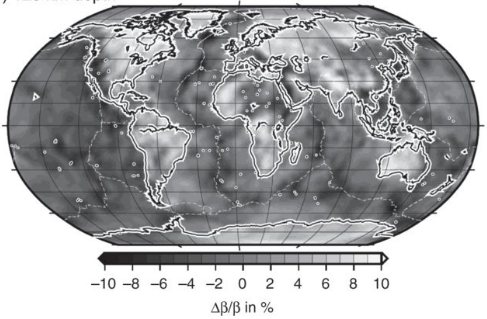

(b) 125 km depth

Fig. 7.41 Horizontal slices through seismic shear velocity structure in the upper mantle. (a) Relative shear velocity variations at 50 km depth. Very low-velocity regions beneath the Andes and the Himalayas actually correspond to anomalously thick crust that extends to more than 50 km depth. (b) Relative shear velocity variations at 125 km depth. The high velocities of continents and the low velocities of mid-ocean ridges are particularly visible (courtesy of Andrew Schaeffer, based on work by Schaeffer and Lebedev, 2013).

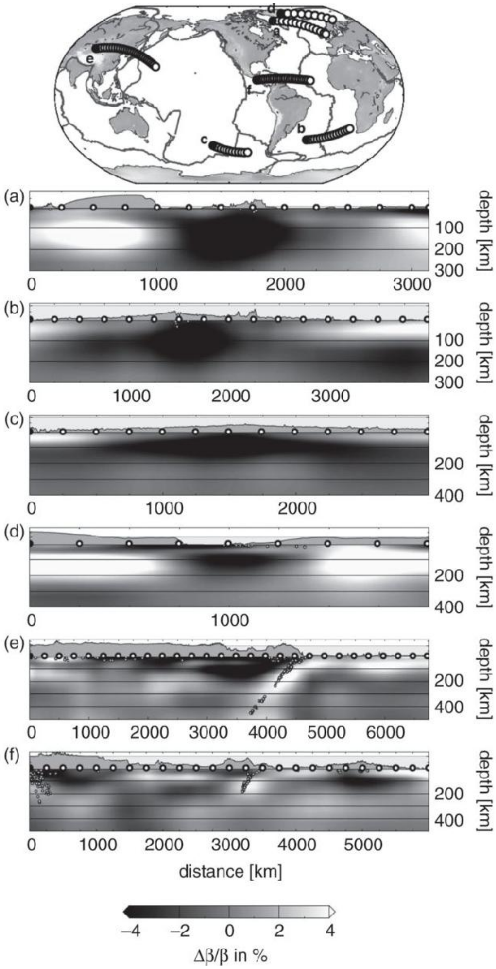


Fig. 7.42 Relative shear velocity variations in the upper mantle. The vertical slices cut through hotspots (Iceland in (a) and Tristan da Cunha in (b), see Fig. 2.21), normal mid-ocean ridges ((c) and (d)), and subduction zones ((e) and (f)). Small dots mark the hypocenters of earthquakes, mostly located in the Wadati–Benioff zones of subducting plates (courtesy of Andrew Schaeffer, based on work by Schaeffer and Lebedev, 2013).

The temperature sensitivity of relative P-wave velocity, $(\Delta\alpha/\alpha)/\Delta T$, is roughly $0.5 \cdot (\Delta\beta/\beta)/\Delta T$, meaning that the P-wave velocity anomaly caused by a certain temperature variation is only about half the S-wave velocity anomaly. Therefore, S-wave tomography tends to be more diagnostic of temperature anomalies in the mantle than P-wave tomography. With increasing depth, both $(\Delta\beta/\beta)/\Delta T$ and $(\Delta\alpha/\alpha)/\Delta T$ decrease, so that temperature variations become less visible. For example, $(\Delta\beta/\beta)/\Delta T$ is around –1% per 100°C at 200 km depth and around –0.5% per 100°C at 600 km depth.

In addition to tracing mid-ocean ridges, narrow low-velocity anomalies of several percent have also been found within the upper mantle beneath most hotspots (Fig. 7.42a,b; Fig. 2.21). Some of these anomalies seem to be confined to the upper 100 km. Others, beneath Iceland and Hawaii, for example, can be followed to several hundred kilometers depth, in agreement with the hypothesis that narrow mantle plumes transport hot material toward the surface (Section 2.8).

The dominant high-velocity anomalies in the upper mantle are found beneath stable continental interiors (Fig. 7.41). Examples include the Canadian shield, the East European Platform, and large parts of central and western Australia. These regions had billions of years to cool since their formation, and the resulting low temperatures are reflected in elevated seismic velocities. The depth extent of the anomalies is typically 150–250 km, which is often interpreted as the thickness of continental lithosphere.

Pronounced high-velocity anomalies are also found beneath subduction zones, where cold oceanic lithosphere descends into the mantle (Fig. 7.42e,f). At some subduction zones, the subducting cold slab is able to penetrate through the transition zone and to sink deeper into the lower mantle. In other cases the 660 km discontinuity appears to block the downward motion of the subducting slab, which is then caused to deflect horizontally. Some slabs that penetrate the 660 km discontinuity appear to sink deep into the lower mantle. Some may in fact reach the core–mantle boundary.

7.4.6.3 The 410 and 660 km Discontinuities

The increase of pressure with increasing depth results in phase transitions toward minerals where atoms in the crystal lattice are more densely packed. At an average depth of approximately 410 km, olivine transforms to spinel, which then transforms to perovskite and ferropericlase around 660 km depth. Seismologically, these transitions are visible as discontinuities where downgoing seismic waves are reflected back to the surface.

The depth of a mantle discontinuity at a specific location may depart from the averages of 410 and 660 km. This is the result of lateral variations in temperature and composition. For example, a temperature decrease of 100°C displaces the 410 km discontinuity around 8 km upwards and the 660 km discontinuity around 6 km downwards. The change in depth (or pressure) of a phase transition with increasing temperature is called the *Clapeyron slope*.

Mapping the depth of the mantle discontinuities is a complicated task because the travel time of waves reflected off a discontinuity depends on its depth and on the seismic velocity between the discontinuity and the surface of the Earth where the seismometer is located. As a consequence, seismic velocity structure and discontinuity depth need to be inferred simultaneously from seismic data.

Despite this complication, the depth variations of the 660 km discontinuity are well known at length scales of several thousand kilometers (Fig. 7.43a). One of the most robust features is an upward deflection by 10–20 km beneath large parts of the Pacific. The corresponding temperature anomaly of approximately 160–320°C is often interpreted as being related to hot upwelling material that may feed numerous volcanoes in the Pacific Ocean, far from plate boundaries (see Section 9.9.4 on superplumes). In contrast, the rims of the Pacific are marked by downwellings of the 660 km discontinuity, which reflect lower than average temperatures. These are most likely related to cold lithospheric slabs that descend into the lower mantle.

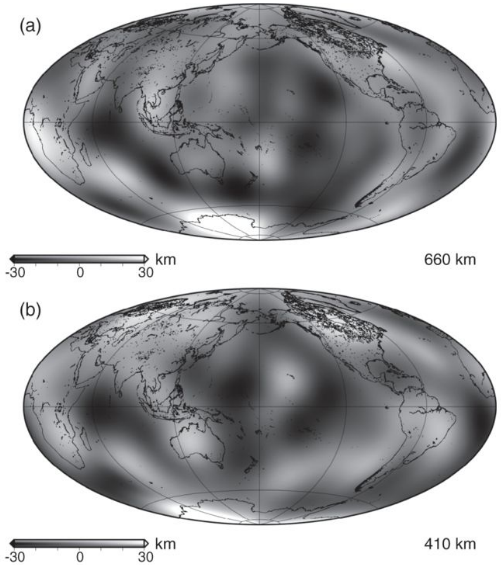


Fig. 7.43 Depth variations of (a) the 660 km discontinuity and (b) the 410 km discontinuity. (Courtesy of Maria Koroni, based on data by Meier et al., 2009.)

Though the 410 km discontinuity is closer to the surface, its depth variations are less well known. This may be due to a more complex structure. In fact, the 410 km discontinuity may not actually be a sharp discontinuity but a rather smooth transition from lower to higher seismic velocities. Among other factors, the width of the transition depends on the amount of hydrogen built into the crystal lattice of olivine, and it may be more than 20 km in some places. Not knowing the thickness of the transition beforehand complicates both the interpretation of the seismic signals and the resulting models of discontinuity depth, such as the one shown in Fig. 7.43b. The strong up- and downward deflections of more than 20 km in the Pacific and Indian Oceans, for example, are the result of variations not only in temperature, but also in chemical composition, including hydrogen content. The relative contributions of these factors remains rather enigmatic.

7.4.6.4 The Lower Mantle

While the seismic velocity structure of the Earth's upper mantle mostly reflects plate tectonic processes, its lower mantle is dominated by two large low-velocity provinces (LLVPs) where both S-wave and P-wave velocities are lower than average (Fig. 7.44). Interestingly, the two LLVPs are nearly antipodal; one is centered below the Pacific, the other at around 180° distance away below Africa. They cover almost half the surface of the Earth's outer core.

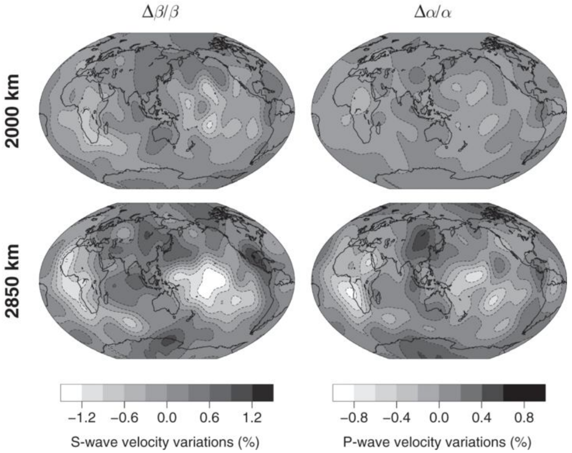


Fig. 7.44 Relative S-wave velocity (left) and P-wave velocity (right) variations in the deep mantle of the Earth at 2000 km depth and at 2850 km depth, just above the core–mantle boundary. (Courtesy of Paula Koelemeijer, based on work by Koelemeijer et al., 2016.)

Despite being well resolved in tomographic images of the lower mantle, the physical nature of the LLVPs remains debated. They may represent anomalously hot material that is being heated from below by the underlying outer core. The elevated temperature would cause the LLVP to be less dense than the surrounding mantle, and therefore buoyant. As a result, the LLVP would slowly rise toward the surface, much like a pocket of hot air rises in the atmosphere. Within this purely thermal interpretation, LLVPs are often referred to as *superplumes*.

Alternatively, LLVPs may acquire their anomalously low velocities by having both elevated temperature and a different chemical composition. The latter may result from the subduction of basaltic oceanic crust, which contains less magnesium and more silicon than normal lower mantle material. Relative to the surrounding lower mantle, the more basaltic material would have a higher density. Thus, in contrast to a purely thermal superplume, this material would not be buoyant. Instead, it would stably reside atop the core–mantle boundary, showing little tendency to rise. In the context of this scenario, LLVPs are frequently termed *thermochemical piles*.

The competing hypotheses have very different implications for the Earth's heat budget. While superplumes would transport heat upwards, thermochemical piles would act as thermal insulators that hinder the transport of heat from the outer core into the mantle. To discriminate between these options, seismologists study the ratio of the relative S-wave velocity variations, $\Delta\beta/\beta$, and the relative P-wave velocity variations, $\Delta\alpha/\alpha$,

$$R = \frac{\Delta\beta/\beta}{\Delta\alpha/\alpha}, \tag{7.28}$$

where $\Delta\beta$ and $\Delta\alpha$ are the absolute velocity variations in m s^{-1}. Mineral physics experiments indicate that values of R below 2–2.5 in the lower mantle are indicative of chemical homogeneity – that is, the absence of material with a different composition. Seismic tomography often finds R slightly above 2.5 between the core–mantle boundary and about 2000 km depth, suggesting the presence of small amounts of chemically distinct material. However, the uncertainties of current tomographic models are still too large to draw final conclusions.

7.5 *Suggestions for Further Reading*

Introductory Level

Bolt, B. A. 2003. *Earthquakes*. New York: W.H. Freeman and Co.

Bryant, E. 2001. *Tsunami: The Underrated Hazard*. Cambridge: Cambridge University Press.

Sykes, L. R. 2019. *Plate Tectonics and Great Earthquakes: 50 Years of Earth-Shattering Events*. New York: Columbia University Press.

Walker, B. S. 1982. *Earthquake*. Alexandria, VA: Time-Life Books Inc.

Intermediate and Advanced Level

Fowler, C. M. R. 2004. *The Solid Earth: An Introduction to Global Geophysics* (2nd ed.). Cambridge: Cambridge University Press.

Igel, H. 2017. *Computational Seismology*. Oxford: Oxford University Press.

Iyer, H. M. and Hirahara, K. (eds.) 1993. *Seismic Tomography: Theory and Practice*. London: Chapman and Hall.

Karato, S.-I. 2013. *Physics and Chemistry of the Deep Earth*. Oxford: Wiley-Blackwell.

Nolet, G. 2008. *A Breviary of Seismic Tomography*. Cambridge: Cambridge University Press.

Science: Special Section. 2005. The Sumatra–Andaman Earthquake. *Science*, **308**, 1125–1146.

7.6 Review Questions

1. How does the elastic-rebound model explain the origin of a tectonic earthquake?

2. What is the epicenter of an earthquake? Explain how the epicenter of an earthquake can be located? What is the minimum number of seismic records needed?

3. Explain the difference between earthquake *intensity* and *magnitude*.

4. If the magnitude of an earthquake is 0.5 greater than that of another, how much greater is the amount of energy it releases?

5. How does a tsunami originate? Why are tsunami barely noticeable over the open ocean but very dangerous near the shore?

6. Describe the geographical distribution of the Earth's seismically active zones.

7. Describe with the aid of sketches the distribution of earthquakes at the three major plate boundaries: (a) spreading ridge, (b) transform fault, and (c) subduction zone.

8. Sketch the fault plane solutions that characterize earthquakes at each type of plate boundary. Describe their significance with regard to plate tectonic motions.

9. How do the distribution and focal solutions of earthquakes at a transform fault differ from those at a strike–slip fault?

10. What is the difference between earthquake prediction, earthquake forecasting, and earthquake early warning?

11. What is the difference between seismic hazard and seismic risk? How do these concepts relate to the intensity and the magnitude of earthquakes?

12. What is the Mohorovičić discontinuity? What is the seismic evidence for this feature? What is its average depth under continents, under oceans, and for the entire world?

13. How do seismic wave velocities change at the major discontinuities in the Earth’s internal structure? How are these discontinuities characterized?

14. On a cross-section of the Earth, sketch the paths of the following seismic rays: (a) PKP, (b) SKS, (c) PcP, (d) PPP.

15. What is a PKIKP wave? Describe the refraction of this wave at each discontinuity it crosses.

16. What are the main features of shear velocity heterogeneities in the upper and the lower mantle of the Earth? How are they related to processes such as plate tectonics?

7.7 Exercises

1. A strong earthquake off the coast of Japan sets off a tsunami that propagates across the Pacific Ocean (average depth $d = 5$ km).

(a) Calculate the velocity of the wave in km h^{-1} and the corresponding wavelength, when the wave has a dominant period of 30 min.

(b) How long does the wave take to reach Hawaii, which is at an angular distance of 54° from the epicenter?

2. The dispersion relation between frequency ω and wave number k of seismic water waves for water depth d is (Box 7.1)

$$\omega^2 = gk \tanh(kd).$$

(a) Modify this expression for wavelengths that are much shorter than the water depth.

(b) Determine the phase velocity of these waves.

(c) Show that the group velocity of the waves is half the phase velocity.

3. In a two-layer Earth the mantle and core are each homogeneous and the radius of the core is one-half the radius of the Earth. Derive a formula for the travel time curve for the arrival time t of the phase PcP at epicentral distance Δ. Verify the formula for the maximum possible value of Δ in this model.

4. The P-wave from an earthquake arrives at a seismic station at 10:20 a.m., and the S-wave arrives at 10:25 a.m. Assuming that the P-wave velocity is 5 km s^{-1} and that Poisson's ratio is 0.25, compute the time at which the earthquake occurred and its epicentral distance in degrees from the station.

5. The following table gives arrival times of P-waves (t_p) and S-waves (t_s) from a nearby earthquake:

Recording station	Time of day [hr:min]	t_p (s)	t_s (s)
A	23:36	54.65	57.90
B	23:36	57.34	62.15
C	23:37	00.49	07.55
D	23:37	01.80	10.00
E	23:37	01.90	10.10
F	23:37	02.25	10.70
G	23:37	03.10	12.00
H	23:37	03.50	12.80
I	23:37	06.08	18.30
J	23:37	07.07	19.79
K	23:37	08.32	21.40
L	23:37	11.12	26.40
M	23:37	11.50	26.20
N	23:37	17.80	37.70

(a) Plot the arrival-time differences ($t_s - t_p$) against the arrival times of the P-wave to produce a Wadati diagram.

(b) Determine the ratio α/β of the seismic velocities.

(c) Determine the time of occurrence (t_0) of the earthquake.

7.8 Computational Exercises

Jupyter notebook computational exercises can be found online for this chapter at www.cambridge.org/FoG3ed (Origin time and location).

8

Geochronology

Preview

The orbital and rotational motions of the Earth and Moon define our familiar units of time. In these units, geological time is immense, involving hundreds of millions, even billions of years. It is subdivided into practical units in terms of the rock and fossil records. These records were used in the infancy of geology to estimate the Earth's age, returning false values because of inherently wrong assumptions. We will see how radiometric dating using the decay schemes of various radioactive isotopes gives consistent, reliable ages, and shows that the Earth is around 4.5 billion years old. The workhorse of modern analyses is the mass spectrometer, which can determine the age of a tiny sample, even of a single grain. Radiometric dating can be used in environmental situations, for example to date great ice-sheets and even groundwater.

8.1 Time

Time is both a philosophical and physical concept. Our awareness of time lies in the ability to determine which of two events occurred before the other. We are conscious of a present in which we live and which replaces continually a past of which we have a memory; we are also conscious of a future, in some aspects predictable, that will replace the present. The progress of time was visualized by Sir Isaac Newton to be like a river that flows involuntarily at a uniform rate. The presumption that time is an independent entity underlies all of classical physics. Although Einstein's *theory of relativity* shows that two observers moving relative to each other will have different perceptions of time, physical phenomena are influenced by this relationship only when velocities approach the speed of light. In everyday usage and in non-relativistic science the Newtonian notion of time as an absolute quantity prevails.

The measurement of time is based on counting cycles (and portions of a cycle) of repetitive phenomena. Prehistoric man distinguished the differences between day and night, he observed the phases of the Moon, and was aware of the regular repetition of the seasons of the year. From these observations, the day, month, and year emerged as the units of time. Only after the development of the clock could the day be subdivided into hours, minutes, and seconds.

8.1.1 The Clock

The earliest clocks were developed by the Egyptians and were later introduced to Greece and from there to Rome. About 2000 BC, the Egyptians invented the water clock (or *clepsydra*). In its primitive form this consisted of a container from which water could escape slowly by a small hole. The progress of time could be measured by observing the change of depth of the water in the container (using graduations on its sides) or by collecting and measuring the amount of water that escaped. The Egyptians (or perhaps the Mesopotamians) are also credited with inventing the sundial. Early sundials consisted of devices – poles, upright stones, pyramids, or obelisks – that cast a shadow; the passage of time was observed by the changing direction and length of the shadow. After trigonometry was developed, dials could be accurately graduated and time precisely measured. Mechanical clocks were invented around AD 1000 but reliably accurate pendulum clocks first came into use in the seventeenth century. Accurate sundials, in which the shadow was cast by a fine wire, were used to check the setting and calibration of mechanical clocks until the nineteenth century.

8.1.2 Units of Time

The day is defined by the rotation of the Earth about its axis. The day can be defined relative to the stars or to the Sun (Fig. 8.1). The time required for the Earth to rotate through 360° about its axis and to return to the same meridian relative to a fixed star defines the *sidereal day*. All sidereal days have the same length. Sidereal time must be used in scientific calculations that require rotational velocity relative to the absolute framework of the stars. The time required for the Earth to rotate about its axis and return to the same meridian relative to the Sun defines the *solar day*. While the Earth is rotating about its axis, it is also moving forward along its orbit. The orbital motion about the Sun covers 360° in about 365 days, so that in one day the Earth moves forward in its orbit by approximately 1°. To return to the solar meridian the Earth must rotate this extra degree. The solar day is therefore slightly longer than the sidereal day. Solar days are not equal in length. For example, at perihelion the Earth is moving faster forward in its orbit than at aphelion (see Fig. 1.4). At perihelion the higher angular rate about the Sun means that the Earth has to rotate through a larger than average angle to catch up with the solar meridian. Thus, at perihelion the solar day is longer than average; at aphelion the opposite is the case. The obliquity of the ecliptic causes a further variation in the length of the solar day. *Mean solar time* is defined in terms of the mean length of the solar day. It is used for most practical purposes on Earth, and is the basis for definition of the hour, minute, and second. One *mean solar day* is equal to exactly 86,400 seconds. The length of the *sidereal day* is approximately 86,164 seconds.

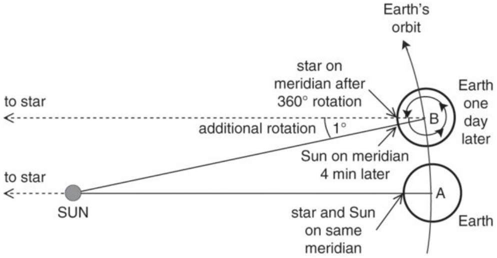

Earth's orbit
star on meridian after 360° rotation
to star
additional rotation
1°
B
Earth one day later
Sun on meridian 4 min later
to star
SUN
A
star and Sun on same meridian
Earth

Fig. 8.1 A sidereal day is the time taken for the Earth to rotate through 360° relative to a fixed star; the solar day is the time taken for a rotation between meridians relative to the Sun. This is slightly more than 360° relative to the stars, because the Earth is also orbiting the Sun.

The *sidereal month* is defined as the time required for the Moon to circle the Earth and return to the celestial longitude of a given star. It is equal to 27.32166 (solar) days. To describe the motion of the Moon relative to the Sun, we have to take into account the Earth's motion around its orbit. The time between successive alignments of the Sun, Earth, and Moon on the same meridian is the *synodic month*. It is equivalent to 29.53059 days.

The *sidereal year* is defined as the time that elapses between successive occupations by the Earth of the same point in its orbit with respect to the stars. It is equal to 365.256 mean solar days. Twice per year, in spring and autumn, the Earth occupies positions in its orbit around the Sun where the lengths of day and night are equal at any point on the Earth. The spring occurrence is called the vernal equinox; that in the autumn is the autumnal equinox. The solar year (correctly called the *tropical year*) is defined as the time between successive vernal equinoxes. It equals 365.242 mean solar days, slightly less than the sidereal year. The small difference (0.014 days, about 20 minutes) is due to the *precession of the equinoxes*, which takes place in the retrograde sense (i.e., opposite to the revolution of the Earth about the Sun) and thereby reduces the length of the tropical year.

Unfortunately the lengths of the sidereal and tropical years are not constant, but change slowly but measurably. In order to have a world standard, the fundamental unit of scientific time was defined in 1956 in terms of the length of the tropical year 1900, which was set equal to 31,556,925.9747 seconds of *ephemeris time*. Even this definition of the second is not constant enough for the needs of modern physics. Highly stable *atomic clocks* have been developed that are capable of exceptional accuracy. For example, the alkali metal cesium has a sharply defined atomic spectral line whose frequency can be determined accurately by resonance with a tuned radio-frequency circuit. This provides the physical definition of the second of *ephemeris time* as the duration of 9,192,631,770 cycles of the cesium atomic clock.

In the mid-nineteenth century, the time at which the Sun reached its zenith (i.e., noon) at Greenwich, England, was adopted as the reference for the measurement of time for civil purposes (such as synchronizing railway timetables). It has been replaced by *universal time* (UT), which is also based on the Earth's rotation. The main form of this time standard is UT1, which uses astronomic and space techniques (e.g., GPS, Section 3.5.4; VLBI, Section 3.5.8) to measure accurately the rotation angle of the Earth. It is the standard measurement of time used for scientific purposes. Coordinated universal time (UTC) is a timescale based on the highly precise atomic clock. It is the basis of civil timekeeping on Earth. UT1 (astronomic) and UTC (atomic) timescales differ slightly, because the Earth's rotation is not completely regular (Section 3.3.5), so occasionally a "leap second" is added or subtracted from UTC to keep the scales within one second of each other. The different ways of defining the basic units of time are important for analyzing some problems in astronomy and satellite geodesy, but for most geophysical applications the minor differences between the different definitions are negligible.

8.1.3 The Geological Timescale

Whereas igneous rocks are formed in discrete, short-lived eruptions of magma, sequences of sedimentary rocks take very long periods of time to form. Many sedimentary formations contain fossils, i.e., relicts of creatures that lived in or near the basin in which the sediments were deposited. Evolution gives a particular identifiable character to the fossils in a given formation, so that it is possible to trace the evolution of different fossil families and correlate their beginnings and extinctions. These characteristics allow correlation of the host rock formation with others that contain part or all of the same fossil assemblage, and permit the development of a scheme for dating sediments relative to other formations. Gradually a *biostratigraphical timescale* has been worked out, which permits the accurate determination of the *relative age* of a sedimentary sequence. This information is intrinsic to any geological timescale.

A geological timescale combines two different types of information. Its basic record is a *chronostratigraphical* scale showing the relationship between rock sequences. These are described in detail in stratigraphical sections that are thought to have continuous sedimentation and to contain complete fossil assemblages. The boundary between two sequences serves as a standard of reference and the section in which it occurs is called a *boundary stratotype*. When ages are associated with the reference points, the scale becomes a geological timescale. Time is divided into major units called *eons*, which are subdivided into *eras*; these in turn are subdivided into *periods* containing several *epochs*. The lengths of time units in early timescales, before ages were determined with radioactive isotopes, were expressed as multiples of the duration of the Eocene epoch. Modern geological timescales are based on isotopic ages, which are calibrated in normal time units.

The geological timescale is constantly being revised and updated. Improved descriptions of important boundaries in global stratotype sections, more refined correlation with other chronostratigraphical units, and better calibration by more accurate isotopic dating lead to frequent revisions. Figure 8.2 shows an example of a current geological timescale.

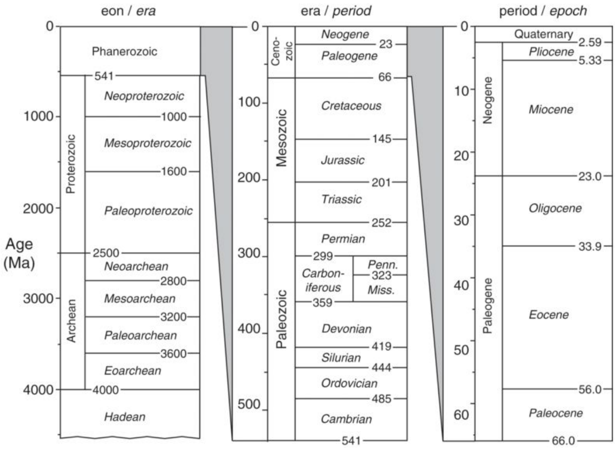


Fig. 8.2 A simplified version of the 2012 geological timescale of the Geological Society of America.

Data source: Gradstein et al., *The Geologic Time Scale* 2012, Amsterdam: Elsevier.

8.2 Historical Estimates of the Earth’s Age

Early estimates of the Earth’s age were dominated by religious beliefs. Some oriental philosophers in antiquity believed that the world had been in existence for millions of years. Yet, Western thought on this topic was dominated for centuries by the tenets of the Jewish and Christian faiths. Biblical estimates of the world’s age were made on the basis of genealogy by adding up the lengths of lifetimes and generations mentioned in the Old Testament. Some estimates also incorporate information from other ancient scriptures or related sources. The computed age of Creation invariably gave an age of the Earth less than 10,000 years.

The best-known biblical estimate of the date of Creation was made by James Ussher (1581–1656), an Irish archbishop. His analysis of events recorded in the Old Testament and contemporary ancient scrolls led Ussher to proclaim that the exact time of Creation was at the beginning of night on October 22, in the year 4004 BC, which makes the Earth approximately 6000 years old. Other biblical scholars inferred similar ages. This type of “creationist” age estimate is still favored by many fundamentalist Christians, whose faith is founded on literal interpretation of the Bible. Biblical scholars with a more broadly based faith recognize age estimates based upon scientific methodology and measurement.

In the late nineteenth century the growth of natural philosophy (as physics was then called) fostered calculations of the Earth’s age from physical properties of the solar system. Among these were estimates based on the cooling of the Sun, the cooling of the Earth, and the slow increase in the Earth–Moon distance. Chemists tried to date the Earth by establishing the time needed for the seas to acquire their salinity, while geologists conjectured how long it would take for sediments and sedimentary rocks to accumulate.

8.2.1 Cooling of the Sun

Setting aside the biblical “sequence” chronicled in the book of Genesis, modern-day philosophers opine that the Earth cannot be older than the Sun. Cooling of the Sun takes place by continuous *radiation* of energy into space. The amount of solar energy falling on a square meter per second at the Earth’s distance from the Sun (1 AU) is called the solar constant; it equals 1360 W m^{-2}. The amount of energy lost by the Sun per second is obtained by multiplying this value by the surface area of a sphere whose radius is 1 AU. This simple calculation shows that the Sun is losing energy at the rate of 3.83×10^{26} W. In the nineteenth century, before the discovery of radioactivity and nuclear reactions, the source of this energy was not known. A German scientist, H. L. F. von Helmholtz, suggested in 1856 that it might result from the change of potential energy due to gravitational condensation of the Sun from an originally more distended body. The condensational energy E_s of a mass M_s of uniform density and radius R_s is given by

$$E_s = \frac{3}{5} G \frac{M_s^2}{R_s} = 2.28 \times 10^{41} \quad \text{J.} \tag{8.1}$$

In this equation G is the gravitational constant (see Section 3.2.1.1). The factor 3/5 in the result arises from the assumption of a uniform density distribution inside the Sun. Dividing the condensational energy by the rate of energy loss gives an age of 19 million years for the Sun. Allowing for the increase of density toward the center causes an approximately threefold increase in the gravitational energy and the inferred age of the Sun.

8.2.2 Cooling of the Earth

In 1862 William Thomson, who later became Lord Kelvin, examined the cooling of the Sun in more detail. Unaware (although suspicious) of other sources of solar energy, he concluded that gravitational condensation could supply the Sun's radiant energy for at least 10 million years, but not longer than 500 million years (500 Myr). In the same year he investigated the cooling history of the Earth. Measurements in deep wells and mines had shown that temperature increases with depth. The rate of increase, or *temperature gradient*, was known to be variable, but seemed to average about 1°F for every 50 ft (0.036°C m^{-1}). Kelvin inferred that the Earth was slowly losing heat and assumed that it did so by the process of *thermal conduction* alone (Section 9.4.1). This enabled him to deduce the Earth's age from a solution of the one-dimensional equation of heat conduction (see Eq. 9.29 and Box 9.2), which relates the temperature T at time t and depth z to physical properties of the cooling body, such as its density (ρ), specific heat (c), and thermal conductivity (k).

These parameters are known for particular rock types, but Kelvin had to adopt generalized values for the entire Earth. The model assumes that the Earth initially had a uniform temperature T_0 throughout, and that it cooled from the outside, leaving the present internal temperatures higher than the surface temperature. After a time t has elapsed the temperature gradient at the surface of the Earth is given by

$$\left(\frac{dT}{dz}\right)_{z=0} = \sqrt{\frac{\rho c}{\pi k}}\frac{T_0}{\sqrt{t}}. \quad (8.2)$$

Kelvin assumed an initial temperature of 7000°F (3871°C; 4144 K) for the hot Earth, and contemporary values for the surface temperature gradient and thermal parameters. The calculation for t yielded an age of about 100 Myr for the cooling Earth.

8.2.3 Increase of the Earth–Moon Separation

The origin of the Moon is still uncertain. George H. Darwin, son of the more famous Charles Darwin and himself a pioneer in tidal theory, speculated that it was torn from the Earth by rapid rotation. Like other classical theories of the Moon's origin (e.g., capture of the Moon from elsewhere in the solar system or accretion in Earth orbit), the theory is flawed. In 1898 Darwin tried to explain the Earth's age from the effects of lunar tidal friction. The tidal bulges on the Earth interact with the Moon's gravitation to produce a decelerating torque that slows down the Earth's rotation and so causes an increase in the length of the day. The equal and opposite reaction is a torque exerted by the Earth on the Moon's orbit that increases its angular momentum. As explained in Section 3.3.5.2, this is achieved by an increase in the distance between the Moon and the Earth and a decrease in the rotation rate of the Moon about the Earth, which increases the length of the month. The Earth's rotation decelerates more rapidly than that of the Moon, so that eventually the angular velocities of the Earth and Moon will be equal. In this synchronous state the day and month will each last about 47 of our present days and the Earth–Moon distance will be about 87 Earth radii; the separation is presently 60.3 Earth radii.

Similar reasoning suggests that earlier in Earth's history, when the Moon was much closer to the Earth, both bodies rotated faster so that an earlier synchronous state may be conjectured. The day and month would each have lasted about five of our present hours and the Earth–Moon distance would have been about 2.3 Earth radii. However, at this distance the Moon would be inside the *Roche limit*, about three Earth radii, at which the Earth's gravitation would tear it apart. Thus, it is unlikely that this condition was ever realized.

Darwin calculated the time needed for the Earth and Moon to progress from an initially unstable close relationship to their present separation and rotation speeds, and concluded that a minimum of

56 Myr would be needed. This provided an independent estimate of the Earth’s age, but it is unfortunately as flawed in its underlying assumptions as other models.

8.2.4 Oceanic Salinity

Several determinations of the age of the Earth have been made on the basis of the chemistry of seawater. The reasoning is that salt is picked up by rivers and transported into lakes and seas; evaporation removes water but the vapor is fresh, so the salt is left behind and accumulates with time. If the accumulation rate is measured, and if the initial salt concentration was zero, the age of the sea, and by inference, the Earth, can be calculated by dividing the present concentration by the accumulation rate.

Different versions of this method have been investigated. The most noted is that of an Irish geologist, John Joly, in 1899. Instead of measuring salt he used the concentration of a pure element, sodium. Joly determined the total amount of sodium in the oceans and the annual amount brought in by rivers. He concluded that the probable maximum age of the Earth was 89 Myr. Later estimates included corrections for possible sodium losses and nonlinear accumulation, but gave similar ages less than about 100 Myr.

The principle flaw in the chemical arguments is the assumption that sodium accumulates continuously in the ocean. In fact, all elements are withdrawn from the ocean at about the same rate as they are brought in. As a result, seawater has a chemically stable composition that is not changing significantly. The chemical methods do not measure the age of the ocean or the Earth, but only the average length of time that sodium resides in the seas before it is removed.

8.2.5 Sedimentary Accumulation

Not to be outdone by the physicists and chemists, late nineteenth-century geologists tried to estimate the Earth's age using stratigraphical evidence for the accumulation of sediments. The first step involved determining the thicknesses of sediment deposited during each unit of geological time. The second step was to find the corresponding sedimentation rate. When these parameters are known, the length of time represented in each unit can be calculated. The age of the Earth is the sum of these times.

The geological estimates are fraught with complications. Sediments with a silicate matrix, such as sandstones and shales, are deposited mechanically, but carbonate rocks form by precipitation from seawater. To calculate the mechanical rate of sedimentation the rate of input has to be known. This requires knowing the area of the depositional basin – the area supplying the sediments – and its rate of land erosion. Rates used in early studies were largely intuitive. The calculations for carbonate rocks required knowing the rate of solution of calcium carbonate from the land surfaces, but could not correct for the variation of solubility with depth in the depositional basin. The number of unknown or crudely known parameters led to numerous divergent geological estimates for the Earth's age, ranging from tens of millions to hundreds of millions of years.

8.3 Radioactivity

In 1896 a French physicist, Henri Becquerel, laid a sample of uranium ore on a wrapped, undeveloped photographic plate. After development the film showed the contour of the sample. The exposure was attributed to invisible rays emitted by the uranium sample. The phenomenon, which became known as *radioactivity*, provides the most reliable methods yet known of dating geological processes and calculating the age of the Earth and solar system. To appreciate radioactivity, we must consider briefly the structure of the atomic nucleus.

The nucleus of an atom contains positively charged protons and electrically neutral neutrons. The electrostatic *Coulomb force* causes the protons to repel each other. It decreases as the inverse square of the separation (see Section 10.2) and so can act over distances that are large compared to the size of the nucleus. An even more powerful *nuclear force* holds the nucleus together. This attractive force acts between protons, between neutrons, and between protons and neutrons. It is only effective at short distances, for a particle separation less than about 3×10^{-15} m.

Suppose that the nucleus of an atom contains *Z* protons and is surrounded by an equal number of negatively charged electrons, so that the atom is electrically neutral; *Z* is the *atomic number* of the element and defines its place in the periodic table. The number *N* of neutrons in the nucleus is its *neutron number*, and the total number *A* of protons and neutrons is the *mass number* of the atom. Atoms of the same element with different neutron numbers are called *isotopes* of the element. For example, uranium contains 92 protons but may have 142, 143, or 146 neutrons. The different isotopes are distinguished by appending the mass number to the chemical symbol, giving ^{234}U, ^{235}U, and ^{238}U.

The Coulomb force of repulsion acts between every pair of protons in a nucleus, while the short-range nuclear force acts only on nearby protons and neutrons. To avoid flying apart due to Coulomb repulsion, all nuclei with atomic number Z greater than about 20 have an excess of neutrons ($N > Z$). This helps to dilute the effects of the repulsion-producing protons. However, nuclei with $Z \geq 83$ are unstable and disintegrate by *radioactive decay*. This means they break up spontaneously by emitting elementary particles and other radiation.

At least 28 distinct elementary particles are known to nuclear physics. The most important correspond to the three types of radiation identified by early investigators and are called α-, β-, and γ-rays. An α-ray (or α-particle) is a helium nucleus that has been stripped of its surrounding electrons; it is made up of two protons and two neutrons and so has atomic number 2 and mass number 4. A β-particle is an electron. Some reactions emit additional energy in the form of γ-rays, which have a very short wavelength and are similar in character to X-rays.

8.3.1 Radioactive Decay

A common type of radioactive decay is when a neutron n_0 in the nucleus of an atom spontaneously changes to a proton, p^+, and a β-particle (non-orbital electron). The β-particle is at once ejected from the nucleus along with another elementary particle called an antineutrino, ν, which has neither mass nor charge and need not concern us further here. The reaction can be written

$$n_0 \Rightarrow p^+ + \beta^- + \nu. \tag{8.3}$$

Radioactive decay is a statistical process. It is customary to call the nucleus that decays the *parent* and the nucleus after decay the *daughter*. It is not possible to say in advance which nucleus will spontaneously decay. But the probability that any one will decay per second is a constant, called the *decay constant* or *decay rate*. The number of disintegrations per second is called the *activity* of the nucleus. The derived unit of radioactivity, corresponding to one disintegration per second, is the *becquerel* (Bq).

Statistical behavior is really only applicable to large numbers, but the process of radioactive decay can be illustrated with a simple example. Suppose we start with 1000 nuclei, and the chance per second of a decay is 1 in 10, or 0.1. In the first second, 10% of the parent nuclei spontaneously decay (i.e., 100 decays take place); the number of parent nuclei is reduced to 900. The probability that any parent decays in the next second is still 1 in 10, so 90 further decays can be expected. Thus, after two seconds the number of parent nuclei is reduced to 810, after three seconds to 729, and so on. The total number of parent nuclei constantly gets smaller but in principle it never reaches zero, although after a long time it approaches zero asymptotically. The decay is described by an *exponential* curve.

If the decay rate is equal to λ, then in a short time interval d*t* the probability that a given nucleus will decay is λdt; if at any time we have P parent nuclei, the number that decay in the following interval dt is $P(\lambda dt)$. The change dP in the number of P parent nuclei in a time interval dt due to spontaneous decays is

$$\begin{aligned} dP &= -\lambda P dt \\ \frac{dP}{dt} &= -\lambda P, \end{aligned} \tag{8.4}$$

which has the solution

$$P = P_0 e^{-\lambda t}. \tag{8.5}$$

Equation (8.5) describes the exponential decay of the number of parent nuclides, starting from an initial number P_0. While the number of parent nuclides diminishes, the number of daughter nuclides D increases (Fig. 8.3). D is the difference between P and P_0 and so is given by

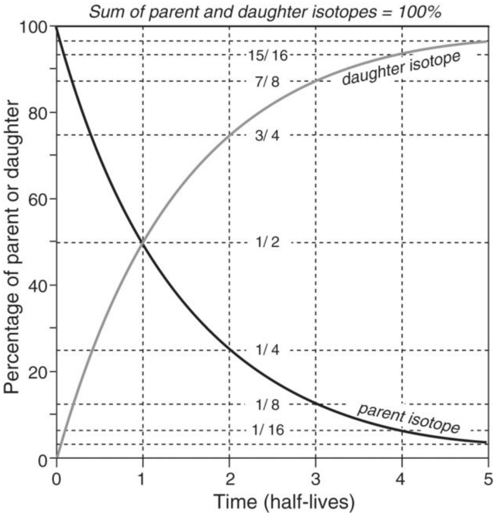


Fig. 8.3 Exponential decrease of the number of parent nuclides and the corresponding growth of the number of daughter nuclides in a typical radioactive decay process.

$$D = P_0 - P = P_0(1 - e^{-\lambda t}). \tag{8.6}$$

The original amount P_0 of the parent nuclide is not known; a rock sample contains a residual amount P of the parent nuclide and an amount D of the daughter product. Eliminating the unknown P_0 from Eq. (8.5) and Eq. (8.6) gives

$$D = P[e^{\lambda t} - 1]. \tag{8.7}$$

Equation (8.4) shows that the number of nuclear disintegrations per second – the activity of the nucleus – is proportional to the number of parent nuclei. The activity A at any given time is thus related to the initial activity A_0 by an equation similar to Eq. (8.5):

$$A = \frac{\partial P}{\partial t} = -\lambda P_0 e^{-\lambda t}$$
$$A = A_0 e^{-\lambda t}. \tag{8.8}$$

The experimental description of radioactive decay by Ernest (later Lord) Rutherford and Frederick Soddy in 1902 was based on the observations of times needed for the activity of radioactive materials to decrease by steps of one-half. This time is known as the *half-life* of the decay. In the first half-life the number of parent nuclides decreases to a half, in the second half-life to a quarter, in the third to an eighth, etc. The number of daughter nuclides increases in like measure, so that the sum of parent and daughter nuclides is always equal to the original number P_0. Letting P/P_0 equal 1/2 in Eq. (8.5), we get the relationship between half-life $t_{1/2}$ and decay constant λ:

$$t_{1/2} = \frac{\ln 2}{\lambda}.$$

(8.9)

The decay rates and half-lives are known for more than 1700 radioactive isotopes. If the half-life of an isotope is longer than the age of the solar system the isotope is said to be *primordial*. Such isotopes are thought to have been formed in processes that occurred before the formation of the solar system. Some isotopes are only produced in nuclear explosions and are so short-lived that they last only a fraction of a second. Other short-lived isotopes are produced by collisions between cosmic rays and atoms in the upper atmosphere and have short half-lives lasting minutes or days.

A number of isotopes are not radioactive and therefore do not decay; they are referred to as *stable isotopes*. Their stability makes them useful as geological and environmental tools. The relative abundance of stable isotopes of the same element (e.g., carbon, nitrogen, or oxygen) is known as a stable isotope ratio. The ratio can be affected by natural processes. Consequently, stable isotopes are useful as environmental indicators. An example of the use of oxygen isotope ratios in sediment and ice cores to reconstruct paleoclimate variations is described in Box 8.1.

Box 8.1 Oxygen Isotope Ratio

The element oxygen has two important stable isotopes. These are a “light” isotope ^{16}O, with 8 protons and 8 neutrons in its nucleus, and a “heavy” isotope ^{18}O with 8 protons and 10 neutrons. Mass spectrometers are capable of determining accurately the masses of these isotopes, and so the mass ratio $^{18}O/^{16}O$ can be measured in quite small samples. In order to describe the deviation of the $^{18}O/^{16}O$ ratio in a given sample from a standard value, a useful parameter, $\delta^{18}O$, is defined as follows:

$$\delta^{18}O = 1000 \times \left(\frac{(^{18}O/^{16}O)_{\text{sample}} - (^{18}O/^{16}O)_{\text{standard}}}{(^{18}O/^{16}O)_{\text{standard}}} \right).$$

The standard value is taken to be the $^{18}O/^{16}O$ ratio in the modern oceans at depths in the 200–500 m range, or alternatively the ratio in a fossil belemnite, known as PDB. The factor 1000 causes $\delta^{18}O$ values to be expressed in parts per thousand. The global climate at present is relatively warmer than in the past, so $^{18}O/^{16}O$ ratios measured in ancient samples tend to be smaller than the standard value and give negative $\delta^{18}O$ values.

The $^{18}O/^{16}O$ ratio in water is dependent on temperature, as evident in the plot of $\delta^{18}O$ in annual precipitation against the site mean temperature at a global set of sites (Fig. B8.1.1). Two factors determine this temperature dependence: cool air holds less moisture than warm air, and the “heavy” isotope ^{18}O condenses more easily than the “light” isotope ^{16}O. From an air mass that moves polewards, precipitation in warmer regions is relatively rich in ^{18}O, whereas in colder regions it is enriched in ^{16}O.

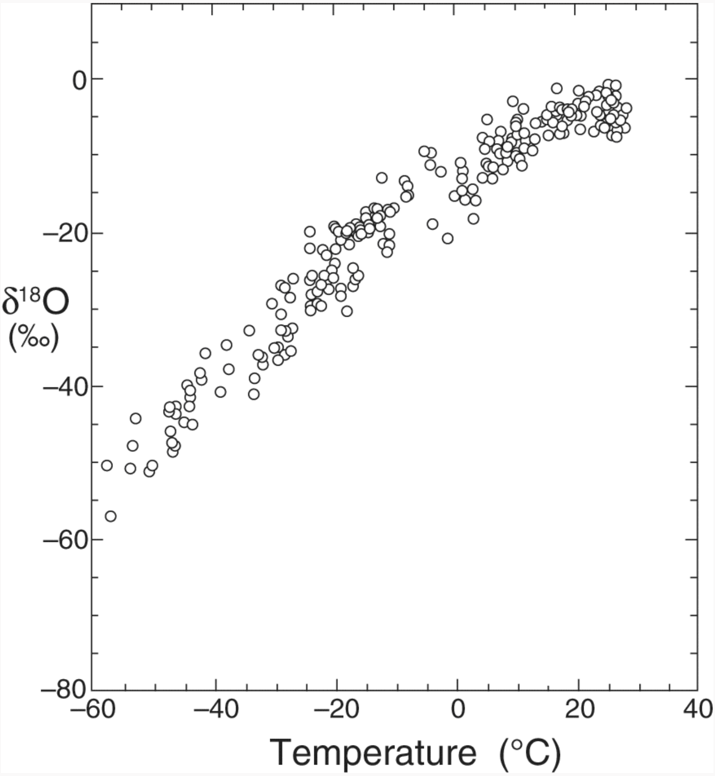


Fig. B8.1.1 Observed correlation with temperature of the oxygen isotope ratio $\delta^{18}O$ measured in precipitation for the present-day climate.

Source: J. Jouzel, R.D. Koster, R.J. Zuozzo, and G.L. Russell, Stable water isotope behavior during the last glacial maximum: A general circulation model analysis, *J. Geophys. Res.*, 99 (D12), 25791–25801, 1994.

The situation is reversed in the oceans. During global warm periods, polar ice-sheets melt, adding fresh water enriched in ^{16}O to the oceans. As a result, low (i.e., large negative) $\delta^{18}O$ values in the oceans indicate intervals of global warming. In contrast, during glacial intervals water is transferred from the oceans to the ice-sheets, primarily as ^{16}O, so the remaining ocean water is enriched in ^{18}O, which increases the $\delta^{18}O$ value. Thus, high (i.e., small negative or even positive) $\delta^{18}O$ values in oceanic water and sediments indicate cool intervals. Conversely, the same conditions result in the opposite relationships between global temperature and $\delta^{18}O$ values measured in polar ice.

Paleoclimatology uses the $\delta^{18}O$ parameter as a guide to past climates, as illustrated by the use of the record in sediment cores from the Ocean Drilling Project to interpret the paleoclimate that affected Chinese loess–paleosol profiles (Fig. 12.11). Samples can be used from a variety of sources: polar ice cores, ocean sediments, and fossil shells. In the last case, biological and chemical processes disturb the simple relationship with temperature, but this can be taken into account and corrected.

A number of naturally occurring isotopes have half-lives of thousands of years (kiloyear, kyr), millions of years (megayear, Myr), or billions of years (gigayear, Gyr), and can be used to determine the ages of geological events. Note that the accepted usage in geochronology is to give durations of time in years (e.g., kyr, Myr, Gyr) and points in time in annum (e.g., ka, Ma, Ga).

8.4 Radiometric Age Determination

Each age-dating scheme involves precise measurement of the concentration of an isotope. This is usually very small. If the radioactive decay has advanced too far, the resolution of the method deteriorates. The best results for a given isotopic decay scheme are obtained for ages less than a few half-lives of the decay. The decay constants and half-lives of some radioactive isotopes commonly used in dating geological events are listed in Table 8.1 and illustrated in Fig. 8.4. Historical and archeological artifacts can be dated with the radioactive carbon method.

Table 8.1 Decay constants and half-lives of some naturally occurring, long-lived radioactive isotopes commonly used in geochronology

Parent isotope	Daughter isotope	Decay constant (10^{-10} yr^{-1})	Half-life (Gyr)
^{40}K	89.5% ^{40}Ca 10.5% ^{40}Ar	5.541	1.251
^{87}Rb	^{87}Sr	0.1408	49.23
^{147}Sm	^{143}Nd	0.0654	106.0
^{232}Th	^{208}Pb	0.4933	14.05
^{235}U	^{207}Pb	9.849	0.7038
^{238}U	^{206}Pb	1.551	4.468

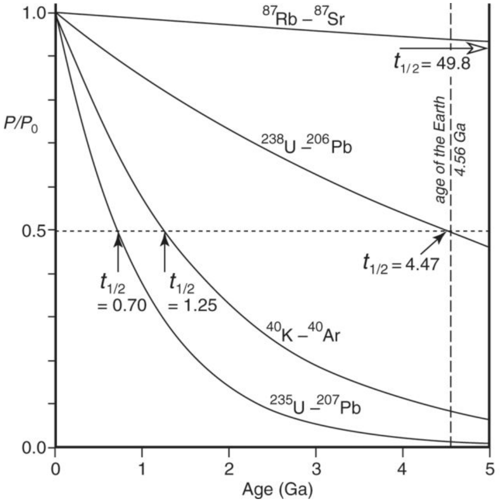


Fig. 8.4 Normalized radioactive decay curves of some important isotopes for dating the Earth and solar system. The arrows indicate the respective half-lives in 10^9 yr.

Data source: G.B. Dalrymple, *The Age of the Earth*, 474 pp., Stanford, CA: Stanford University Press, 1991.

8.4.1 Radioactive Carbon

The Earth is constantly being bombarded by cosmic radiation from outer space. Collisions of cosmic particles with atoms of oxygen and nitrogen in the Earth's atmosphere produce high-energy neutrons. These in turn collide with nitrogen nuclei, transforming them into ^{14}C, a radioactive isotope of carbon. ^{14}C decays by β-particle emission to ^{14}N with a decay rate of ($\lambda = 1.2097 \times 10^{-4}$ yr^{-1}) and a half-life of 5730 years. The production of new ^{14}C is balanced by the loss due to decay, so that a natural equilibrium exists. Photosynthesis in plants replenishes living tissue using carbon dioxide, which contains a steady proportion of ^{14}C. When an organism dies, the renewal stops and the residual ^{14}C in the organism decays radioactively.

The method of dating with radioactive carbon (also referred to as radiocarbon dating, or ^{14}C dating) is a simple decay analysis based on Eq. (8.5). The remaining proportion P of ^{14}C is measured by counting the current rate of β-particle activity, which is proportional to P. This is compared to the original equilibrium concentration P_0. The time since the onset of decay is calculated by solving Eq. (8.5) using the known decay rate for ^{14}C.

The radiocarbon method has been valuable in dating events in the Holocene epoch, which covers the last 10,000 years of geological time, as well as events related to human prehistory. Unfortunately, human activity has disturbed the natural equilibrium of the replenishment and decay scheme. The concentration of ^{14}C in atmospheric carbon has changed dramatically since the start of the industrial age. Partly this is due to the combustion of fossil fuels like coal and oil as energy sources; they have long lost any ^{14}C and dilute its presence in atmospheric carbon. In the past half-century, atmospheric testing of nuclear weapons doubled the concentration of ^{14}C in the atmosphere. Older materials may still be dated by the ^{14}C method, although natural fluctuations in P_0 must be taken into account. These are due to variations in intensity of the geomagnetic field, which acts as a partial shield against cosmic radiation.

8.4.2 The Mass Spectrometer

In the years before World War II, physicists invented the mass spectrometer, an instrument for measuring the mass of an ion. The instrument was further refined during the development of the atomic bomb. After the war it was adopted into the earth sciences to determine isotopic ratios and became a vital part of the process of isotopic age determination.

The mass spectrometer (Fig. 8.5) utilizes the different effects of electrical and magnetic fields on a charged particle, or ion. First, the element of interest is extracted from selected mineral fractions or from pulverized whole rock. The extract is purified chemically before being introduced into the mass spectrometer, where it is ionized. For the analysis of a gas like argon, bombardment by a stream of electrons may be used. Solid elements, such as potassium, rubidium, strontium, or uranium, are vaporized by depositing a sample on an electrically heated filament, or by heating with a high-energy laser beam so that the sample is vaporized directly. The latter process is known as “laser ablation.” The ions then enter an evacuated chamber and pass through an “ion gun,” where they are accelerated by an electrical field and filtered by a velocity selector. This device uses electrical and magnetic fields at right angles to the ion beam to allow only the passage of ions with a selected velocity v. The ion beam is next subjected to a powerful uniform magnetic field B at right angles to its direction of motion. An ion with charge q experiences a Lorentz force (see Box 11.2) equal to (qvB) perpendicular to its velocity and to the magnetic field. Its trajectory is bent to form a circular arc of radius r; the centrifugal force on the particle is equal to (mv^2/r). The curved path focuses the beam on a collector device that measures the intensity of the incident beam. The radius of the circular arc is given by equating the Lorentz and centrifugal forces:

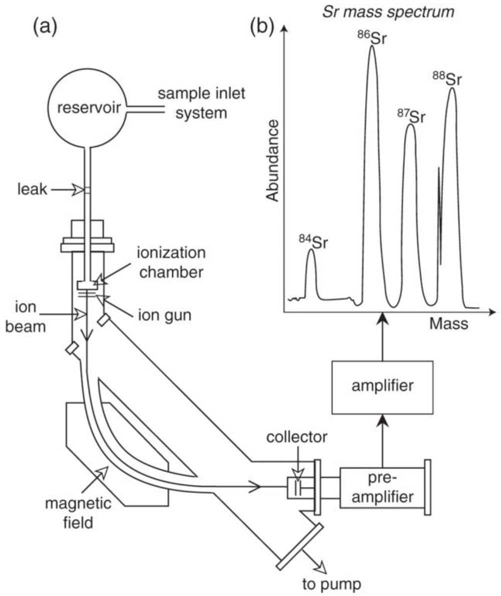

(a)
reservoir
sample inlet system
leak
ionization chamber
ion beam
ion gun
magnetic field
collector
pre-amplifier
to pump
(b)
Sr mass spectrum
Abundance
^{86}Sr
^{88}Sr
^{87}Sr
^{84}Sr
Mass
amplifier

Fig. 8.5 (a) Schematic design of a mass spectrometer; and (b) hypothetical mass spectrum for analysis of Sr isotopes.

Source: D. York and R.M. Farquhar, *The Earth's Age and Geochronology*, 178 pp., Oxford: Pergamon Press, 1972.

$$r = \frac{m}{B}\frac{v}{q}. \tag{8.10}$$

The focal point of the path is determined by the mass of the ion and the strength of the magnetic field *B*. In the case of a strontium analysis, the ion beam leaving the ion gun contains the four isotopes ^{88}Sr, ^{87}Sr, ^{86}Sr, and ^{84}Sr. The beam splits along four paths, each with a different curvature. The magnetic field is adjusted so that only one isotope at a time falls on the collector. The incident current is amplified electronically and recorded. A spectrum is obtained with peaks corresponding to the incidence of individual isotopes (Fig. 8.5b). The intensity of each peak is proportional to the abundance of the isotope, which can be measured with a precision of about 0.1%. However, the relative peak heights give the relative abundances of the isotopes to a precision of about 0.001%.

Significant improvements in the design of mass spectrometer systems have increased their sensitivity further. Following laser ablation, the vaporized sub-micron-sized particles are mixed with an argon plasma. This is a gas consisting of positively charged argon ions and an equal number of unbound electrons at a temperature of 6000–10,000 K. The plasma ionizes the particles, which are extracted into a high vacuum before passing through an ion gun into the mass spectrometer. A drawback of the single-collector design described above is that it is vulnerable to fluctuations in the beam intensity. This problem has been overcome with the introduction of multiple-collector instruments, which allow simultaneous measurement of the split isotopic beams. The *inductively coupled plasma mass spectrometer* (ICP-MS) is a highly sensitive, versatile instrument for determining trace element abundances with a detection level in the region of parts per trillion. It can be used for analyzing both liquid and solid samples, and has many applications in fields outside the earth and environmental sciences.

An important development in the field of mass spectrometry is the ion microprobe mass analyzer. In conventional mass spectrometry, the analysis of a particular element is preceded by separating it chemically from a rock sample. The study of individual minerals or the variation of isotopic composition in a grain is very difficult. The ion microprobe avoids contamination problems that can arise during the chemical separation, and its high resolution permits isotopic analysis of very small volumes. Before a sample is examined in the instrument its surface is coated with gold or carbon. A narrow beam of negatively charged oxygen ions, about 3–10 μm wide, is focused on a selected grain. Ions are sputtered out of the surface of the mineral grain by the impacting ion beam, accelerated by an electrical field, and separated magnetically as in a conventional mass spectrometer. The ion microprobe allows description of isotopic concentrations and distributions in the surface layer of the grain and provides isotopic ratios. The instrument allows isotopic dating of individual mineral grains in a rock, and is especially well suited to the analysis of very small samples.

8.4.3 Rubidium–Strontium

The use of a radioactive decay scheme as given by Eq. (8.7) assumes that the amount of daughter isotope in a sample has been created only by the decay of a parent isotope in a closed system. Usually, however, an unknown initial amount of the daughter isotope is present, so that the amount measured is the sum of the initial concentration D_0 and the fraction derived from decay of the parent P_0. The decay equation is modified to

$$D = D_0 + P(e^{\lambda t} - 1). \tag{8.11}$$

The need to know the amount of initial daughter isotope D_0 is obviated by the analytical method, which makes use of a third isotope of the daughter element to *normalize* the concentrations of daughter and parent isotopes. The rubidium–strontium method illustrates this technique.

Radioactive rubidium (^{87}Rb) decays by β-particle emission to radiogenic strontium (^{87}Sr) with a decay constant $\lambda = 1.393 \times 10^{-11}$ yr^{-1}. The non-radiogenic, stable isotope ^{86}Sr has approximately the same abundance as the radiogenic product ^{87}Sr, and is chosen for normalization. Writing ^{87}Rb for P, ^{87}Sr for D in Eq. (8.11) and dividing both sides by ^{86}Sr gives

$$\left(\frac{^{87}\mathrm{Sr}}{^{86}\mathrm{Sr}}\right) = \left(\frac{^{87}\mathrm{Sr}}{^{86}\mathrm{Sr}}\right)_0 + \left(\frac{^{87}\mathrm{Rb}}{^{86}\mathrm{Sr}}\right)(e^{\lambda t} - 1). \tag{8.12}$$

In a magmatic rock, the isotopic ratio (^{87}Sr/^{86}Sr)$_0$ is uniform in all minerals precipitated from the melt because the isotopes are chemically identical (i.e., they have the same atomic number). However, the proportion of the different elements Rb and Sr varies from one mineral to another. Eq. (8.12) can be compared with the equation of a straight line, such as

$$y = y_0 + mx. \tag{8.13}$$

The ratio $^{87}\mathrm{Sr}/^{86}\mathrm{Sr}$ is the dependent variable, *y*, and the ratio ($^{87}\mathrm{Rb}/^{86}\mathrm{Sr}$) is the independent variable, *x*. If we measure the isotopic ratios in several samples of the rock and plot the ratio $^{87}\mathrm{Sr}/^{86}\mathrm{Sr}$ against the ratio $^{87}\mathrm{Rb}/^{86}\mathrm{Sr}$, we get a straight line, called an *isochron* (Fig. 8.6). The intercept with the ordinate axis gives the initial ratio of the daughter isotope. The slope (*m*) of the line gives the age of the rock, using the decay constant λ:

$$t = \frac{1}{\lambda}\ln(1+m) = 7.042 \times 10^{10}\ln(1+m). \tag{8.14}$$

Because of its long half-life of 49.76 Gyr (Fig. 8.4) the Rb–Sr method is suitable for dating very old events in Earth's history. It has been used to obtain the ages of meteorites and lunar samples, as well as some of the oldest rocks on Earth. For example, the slope of the Rb–Sr isochron in Fig. 8.6 yields an age of 3.62 Ga for the Early Precambrian (Archean) Uivak gneisses from eastern Labrador.

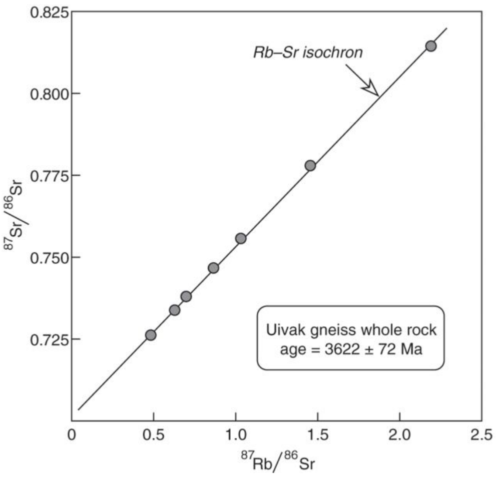


Fig. 8.6 Rb–Sr isochron for the Uivak gneisses from eastern Labrador. The slope of the isochron gives an age of 3.622 ± 0.072 G.

Source: R.W. Hurst, D. Bridgwater, K.D. Collerson and G.W. Weatherill, 3600 m.y. Rb–Sr ages from very early Archean gneisses from Saglek Bay, Labrador, *Earth Planet. Sci. Lett.* 27, 393–403, 1975.

The Rb–Sr and other methods of isotopic dating can be applied to whole rock samples or to individual minerals separated from the rock. The decay equation applies only to a closed system, i.e., to rocks or minerals that have undergone no loss or addition of the parent or daughter isotope since they formed. A change is more likely in a small mineral grain than in the rock as a whole. Individual mineral isochrons may express the age of postformational metamorphism, while the whole rock isochron gives an older age. On the other hand, spurious ages may result from whole rock analyses if samples of different origin, and hence differing composition, are inadvertently used to construct an isochron.

8.4.4 Potassium–Argon

For several reasons the potassium–argon (K–Ar) method is probably the age-dating technique most commonly used by geologists. The parent isotope, potassium, is common in rocks and minerals, while the daughter isotope, argon, is an inert gas that does not combine with other elements. The half-life of 1251 Myr (1.251 Gyr) is very convenient. On the one hand, the Earth's age is equal to only a few half-lives, so radiogenic ^{40}K is still present in the oldest rocks; on the other hand, enough of the daughter isotope ^{40}Ar accumulates in about 10^4 years to give fine resolution. In the late 1950s the sensitivity of mass spectrometers was improved by constructing instruments that could be pre-heated to high temperature to drive off contaminating atmospheric argon. This made it possible to use the K–Ar method for dating lavas as young as a few million years.

Radioactive ^{40}K constitutes only 0.01167% of the K in rocks. It decays in two different ways: (1) by β-particle emission to $^{40}Ca_{20}$ with decay rate $\lambda_{Ca} = 4.962 \times 10^{-10}\ yr^{-1}$; and (2) by electron capture to $^{40}Ar_{18}$ with decay rate $\lambda_{Ar} = 0.581 \times 10^{-10}\ yr^{-1}$. The combined decay constant ($\lambda = \lambda_{Ca} + \lambda_{Ar}$) is equal to $5.543 \times 10^{-10}\ yr^{-1}$. The decay schemes are, respectively:

$$\begin{array}{lllll} (a) & {}^{40}K_{19} & \Rightarrow & {}^{40}Ca_{20} + \beta^- \\ (b) & {}^{40}K_{19} + e & \Rightarrow & {}^{40}Ar_{18}. \end{array} \quad (8.15)$$

Electron capture by a nucleus is more difficult and rarer than β-particle emission, so the decay of $^{40}K_{19}$ to $^{40}Ca_{20}$ is more common than the formation of $^{40}Ar_{18}$. The ratio of electron capture to β-particle decay, $\lambda_{Ar}/\lambda_{Ca}$, is called the *branching ratio*; it equals 0.117. Thus, only the fraction $\lambda_{Ar}/(\lambda_{Ar} + \lambda_{Ca})$, or 10.5%, of the initial radioactive potassium decays to argon. The initial amount of radiogenic ^{40}Ca usually cannot be determined, so the decay to Ca is not used. Allowing for the branching ratio, the K–Ar decay equation is analogous to Eq. (8.7), with ^{40}Ar for the accumulated amount of the daughter product D and ^{40}K for the residual amount of the parent product P:

$$
\begin{aligned}
{}^{40}\mathrm{Ar} &= \frac{\lambda_{\mathrm{Ar}}}{\lambda_{\mathrm{Ar}}+\lambda_{\mathrm{Ca}}}\,{}^{40}\mathrm{K}\left(e^{\lambda t}-1\right) \\
&= 0.1048\ {}^{40}\mathrm{K}\left(e^{\lambda t}-1\right).
\end{aligned}
\tag{8.16}
$$

The potassium–argon method is an exception to the need to use isochrons. It is sometimes called an accumulation clock, because it is based on the amount of ^{40}Ar that has accumulated. It involves separate measurements of the concentrations of the parent and daughter isotopes. The amount of ^{40}K is a small but constant fraction (0.01167%) of the total amount of K, which can be measured chemically. The ^{40}Ar is determined by mixing with a known amount of another isotope, ^{38}Ar, before being introduced into the mass spectrometer. The *relative* abundance of the two argon isotopes is measured and the concentration of ^{40}Ar is found using the known amount of ^{38}Ar. By reordering Eq. (8.16) and substituting for the decay constant, the age of the rock is obtained from the K–Ar age equation:

$$
\begin{aligned}
e^{\lambda t} &= 1+9.54\,\frac{^{40}\mathrm{Ar}}{^{40}\mathrm{K}} \\
t &= 1.804\times 10^{9}\ \ln\left(1+9.54\,\frac{^{40}\mathrm{Ar}}{^{40}\mathrm{K}}\right).
\end{aligned}
\tag{8.17}
$$

The ^{40}Ar in a molten rock easily escapes from the melt. It may be assumed that all of the radiogenic ^{40}Ar now present in a rock has formed and accumulated since the solidification of the rock. The method works well on igneous rocks that have not been heated since they formed. It cannot be used in sedimentary rocks that consist of the detritus of older rocks. Often, it is unsuccessful in metamorphic rocks, which may have complicated thermal histories. A heating phase may drive out the argon, thereby resetting the accumulation clock. This problem limits the usefulness of the K–Ar method for dating meteorites (which have a fiery entry into the Earth’s atmosphere) and very old terrestrial rocks (because of their unknown thermal histories). The K–Ar method can be used for dating lunar basalts, as they have not been reheated since their formation.

8.4.5 Argon–Argon

Some uncertainties related to postformational heating of a rock are overcome in a modification of the K–Ar method that uses the $^{40}Ar/^{39}Ar$ isotopic ratio. The method requires conversion of the ^{39}K in the rock to ^{39}Ar. This is achieved by irradiating the sample with fast neutrons in an atomic reactor.

The terms *slow* and *fast* refer to the energy of the neutron radiation. The energy of slow neutrons is comparable to their thermal energy at room temperature; they are also referred to as *thermal* neutrons. Slow neutrons can be captured and incorporated into a nucleus, changing its size without altering its atomic number. The capture of slow neutrons can increase the size of unstable uranium nuclei beyond a critical value and initiate fission. In contrast, fast neutrons act on a nucleus like projectiles. When a fast neutron collides with a nucleus it may eject a neutron or proton, while itself being captured. If the ejected particle is another neutron, no effective change results. But if the ejected particle is a proton (with a β-particle to conserve charge), the atomic number of the nucleus is changed. For example, bombarding ^{39}K nuclei in a rock sample with fast neutrons converts a fraction of them to ^{39}Ar.

To determine this fraction, a control sample of known age is irradiated at the same time. By monitoring the change in isotopic ratios in the control sample, the fraction of ^{39}K nuclei converted to ^{39}Ar can be deduced. The age equation is similar to that given in Eq. (8.17) for the K–Ar method. However, ^{39}Ar replaces ^{40}K and an empirical constant J replaces the constant 9.54. The value of J is found from the control sample whose age is known. The $^{40}Ar/^{39}Ar$ age equation is

$$t = 1.804 \times 10^9 \ \ln\left(1 + J\,\frac{^{40}\mathrm{Ar}}{^{39}\mathrm{Ar}}\right). \tag{8.18}$$

In the ^{40}Ar/^{39}Ar method the sample is heated progressively to drive out argon at successively higher temperatures. The ^{40}Ar/^{39}Ar isotopic ratio of the argon released at each temperature is determined in a mass spectrometer. The age computed for each increment is plotted against the percentage of Ar released. This yields an *age spectrum*. If the rock has not been heated since it was formed, the argon increments given out at each heating stage will yield the same age (Fig. 8.7a). An isochron can be constructed as in the Rb–Sr method by measuring the abundance of a non-radiogenic ^{36}Ar fraction and comparing the isotopic ratios ^{40}Ar/^{36}Ar and ^{39}Ar/^{36}Ar. In an unheated sample, all points fall on the same straight line (Fig. 8.7b).

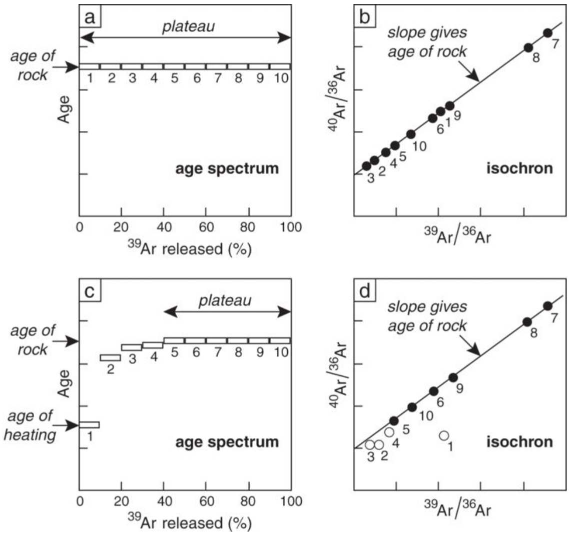


Fig. 8.7 (a) Hypothetical age spectrum and (b) ^{40}Ar/^{39}Ar isochron for a sample that has experienced no secondary heating; (c) hypothetical age spectrum and (d) ^{40}Ar/^{39}Ar isochron for a sample that was reheated but retains some original argon. Numbers on the plots indicate progressive heating stages.

Source: G.B. Dalrymple, *The Age of the Earth*, 474 pp., Stanford, CA: Stanford University Press, Stanford, 1991

If the rock has undergone postformational heating, the argon formed since the heating is released at lower temperatures than the original argon (Fig. 8.7c). It is not known for certain why this is the case. The argon probably passes out of the solid rock by diffusion, which is a thermally activated process that depends on both the temperature and the duration of heating. Unless the postformational heating is long and thorough, only the outside parts of grains may be forced to release trapped argon, while the Ar located in deeper regions of the grains is retained. In a sample that still contains original argon the results obtained at high temperatures form a *plateau* of consistent ages (Fig. 8.7c) from which the optimum age and its uncertainty may be calculated. On an isochron diagram reheated points deviate from the straight line defined by the high-temperature isotopic ratio (Fig. 8.7d).

The high precision of $^{40}Ar/^{39}Ar$ dating is demonstrated by the analysis of samples of melt rock from the Chicxulub impact crater in the Yucatan peninsula of Mexico. The crater is the favored candidate for the impact site of a 10 km diameter meteorite that caused global extinctions at the close of the Cretaceous period. Laser heating was used to release the argon isotopes. The age spectra of three samples show no argon loss (Fig. 8.8). The plateau ages are precisely defined and give a weighted mean age of 64.98 ± 0.05 Ma for the impact. This agrees closely with the 65.01 ± 0.08 Ma age of tektite glass found at a Cretaceous–Tertiary boundary section in Haiti, and so ties the impact age to the faunal extinctions.

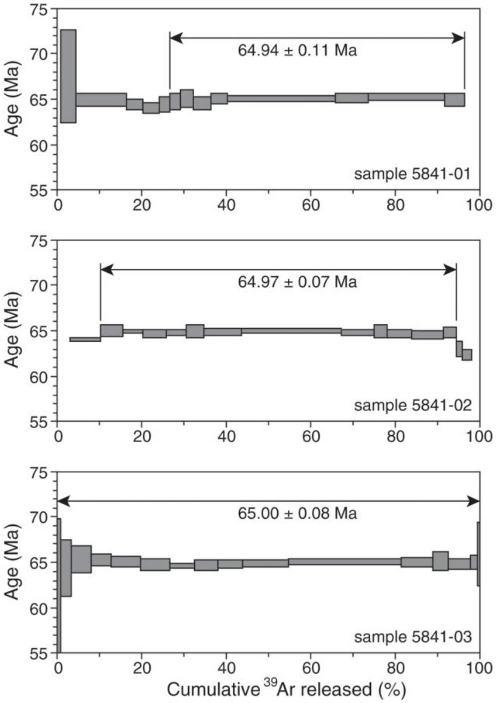


Fig. 8.8 ^{40}Ar/^{39}Ar age spectra for three samples of melt rocks from the Chicxulub Cretaceous–Tertiary impact crater.

Source: C.C. Swisher III, et al., Coeval 40Ar/39Ar ages of 65.0 million years ago from Chicxulub crater melt rock and Cretaceous–Tertiary boundary tektites, *Science* 257, 954–958, 1992.

The age spectrum obtained in Ar–Ar dating of the Menow meteorite is more complicated and shows the effects of ancient argon loss. Different ages are obtained during incremental heating below about 1200°C (Fig. 8.9). The plateau ages above this temperature indicate a mean age of 4.48 ± 0.06 Ga. The shape of the age spectrum suggests that about 25% of the Ar was lost much later, about 2.5 Gyr ago.

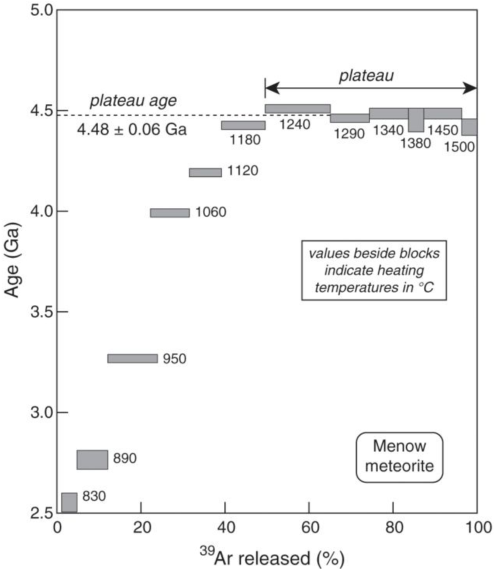


Fig. 8.9 ^{40}Ar/^{39}Ar age spectrum for the Menow meteorite, showing a plateau at 4.48 ± 0.06 Ga.

Data source: G. Turner, M.C. Enright and P.H. Cadogan, The early history of chondrite parent bodies inferred from 40Ar–39Ar ages, in: *Proceedings of the Ninth Lunar and Planetary Science Conference*, pp. 989–1025, 1978.

8.4.6 Uranium–Lead: the Concordia–Discordia Diagram

Uranium isotopes decay through a series of intermediate radioactive daughter products, but eventually they result in stable end-product isotopes of lead. Each of the decays is a multi-stage process but can be described as though it had a single decay constant. We can describe the decay of ^{238}U to ^{206}Pb by

$$\frac{^{206}\mathrm{Pb}}{^{238}\mathrm{U}} = e^{\lambda_{238} t} - 1. \tag{8.19}$$

Likewise, the decay of ^{235}U to ^{207}Pb can be written

$$\frac{^{207}\mathrm{Pb}}{^{235}\mathrm{U}} = e^{\lambda_{235} t} - 1. \tag{8.20}$$

The ^{235}U and ^{238}U isotopes have well-known decay constants: $\lambda_{235} = 9.849 \times 10^{-10}\ \mathrm{yr}^{-1}$, $\lambda_{238} = 1.551 \times 10^{-10}\ \mathrm{yr}^{-1}$. A graph of the ^{206}Pb/^{238}U ratio against the ^{207}Pb/^{235}U ratio is a curve, called the *concordia* line (Fig. 8.10). The abscissa and ordinate of each age on this curve are calculated with Eqs. (8.19) and (8.20), respectively. The concordia line has a curved shape because the decay rates of the uranium isotopes, λ_{238} and λ_{235} respectively, are different.

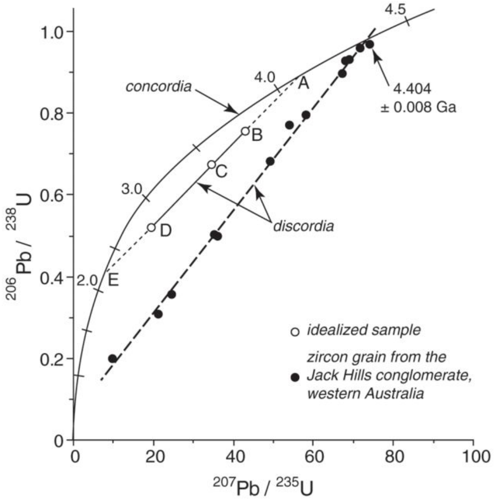


Fig. 8.10 U–Pb concordia–discordia diagram. Lead loss gives points B, C, and D on a hypothetical discordia line. It intersects the concordia curve at A, the age of the rock, and at E, the time of the event that caused lead loss. Marks on the concordia curve indicate age in Ga. The lower discordia line shows actual measurements made on a single zircon grain from the Jack Hills conglomerate in western Australia. The 4.404 ± 0.008 Ga age is the oldest measured for a terrestrial mineral.

Data source: Wilde, S.A., et al., Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Gyr ago, *Nature*, 409, 175–178, 2001.

The amounts of the daughter lead isotopes accumulate at different rates. Lead is a volatile element and is easily lost from minerals, but this does not alter the isotopic ratio of the lead that remains. Loss of lead will cause a point to deviate from the concordia line. However, because the isotopic ratio remains constant the deviant point lies on a straight line between the original age and the time of the Pb-loss. This line corresponds to points that do not agree with the concordia curve; it is called the *discordia* line A hypothetical example is shown in Fig. 8.10. Different mineral grains in the same rock experience different amounts of lead loss, so the isotopic ratios in these grains give different points B, C, and D on the discordia line. The intersection of the concordia and discordia lines at A gives the original age of the rock, or in the case that it has lost all of its original lead, the age of this event. The intersection of the lines at E gives the age of the event that caused the lead loss.

The U–Pb method has been used to date some of the oldest rocks on Earth. The lower discordia line in Fig. 8.10 is a plot of U–Pb measurements made on a tiny zircon grain from the Jack Hills conglomerate in western Australia. The age of lead loss in this grain could not be resolved. The age at which the discordia line intercepts the concordia curve is around 4.36 Ga. A single measurement on a part of the zircon crystal falls very close to the concordia curve and gives an age of 4.404 ± 0.008 Ga. This is the oldest age yet determined on a terrestrial mineral.

8.4.7 Lead–Lead Isochrons

It is possible to construct isochron diagrams for the U–Pb decay systems. This is done (as in the Rb/Sr decay scheme) by expressing the radiogenic isotopes ^{206}Pb and ^{207}Pb as ratios of ^{204}Pb, the non-radiogenic isotope of lead. The decay of ^{238}U to ^{206}Pb (or of ^{235}U to ^{207}Pb) can then be described as in Eq. (8.11), where the initial value of the daughter product is unknown. This gives the decay equations

$$\left(\frac{^{207}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)-\left(\frac{^{207}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)_0=\left(\frac{^{235}\mathrm{U}}{^{204}\mathrm{Pb}}\right)\left(e^{\lambda_{235}t}-1\right), \tag{8.21}$$

$$\left(\frac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)-\left(\frac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)_0=\left(\frac{^{238}\mathrm{U}}{^{204}\mathrm{Pb}}\right)\left(e^{\lambda_{238}t}-1\right). \tag{8.22}$$

These equations can be combined into a single isochron equation:

$$\frac{\left(\frac{^{207}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)-\left(\frac{^{207}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)_0}{\left(\frac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)-\left(\frac{^{206}\mathrm{Pb}}{^{204}\mathrm{Pb}}\right)_0}=\left(\frac{^{235}\mathrm{U}}{^{238}\mathrm{U}}\right)\frac{\left(e^{\lambda_{235}t}-1\right)}{\left(e^{\lambda_{238}t}-1\right)}. \tag{8.23}$$

The ratio of ^{235}U to ^{238}U as measured today has been found to have a constant value 1/137.88 in lunar and terrestrial rocks and in meteorites. The decay constants are well known, so for a given age *t* the right side in Eq. (8.23) is constant. The equation then has the form

$$\frac{y-y_0}{x-x_0}=m. \tag{8.24}$$

This is the equation of a straight line of slope *m* through the point

(x_0, y_0). The initial values of the lead isotope ratios are not known, but a plot of the isotopic ratio $^{207}\mathrm{Pb}/^{204}\mathrm{Pb}$ against the ratio $^{206}\mathrm{Pb}/^{204}\mathrm{Pb}$ is a straight line. The age of the rock cannot be found algebraically. Values for t must be inserted successively on the right side of Eq. (8.23) until the observed slope is obtained.

8.4.8 Radiometric Dating of Environmental Processes

In order to determine the ages of ancient geological events, radioactive decay schemes with very long half-lives must be used. Shorter-lived isotopes can be used to date current and geologically more recent processes that are environmentally important. Numerous radioactive isotopes can be used for this purpose. The best known is radioactive carbon, ^{14}C (Section 8.4.1). When possible anthropological contamination can be taken into account, the radiocarbon half-life of 5730 years allows the dating of events in human prehistory with ages in the range 1000–50,000 years. Other radioactive isotopes with longer (and shorter) half-lives can also be used for environmental studies.

Primary cosmic rays consist mainly of protons and alpha particles that originated outside the solar system (Section 10.3). They collide with the nitrogen and oxygen in the Earth's atmosphere to form a number of radioactive isotopes, which, due to their mode of formation, are called cosmogenic nuclides. Collisions between cosmic rays and atmospheric gases result in a cascade of secondary particles. An important cosmogenic isotope of beryllium, ^{10}Be, is produced in the atmosphere when cosmic rays collide with oxygen (^{16}O). The ^{10}Be decays to stable boron with a half-life of 1.387×10^6 years, which is long enough to make it useful for dating geological records that are older than the 50 kyr datable with ^{14}C. The ^{10}Be dissolves quickly in rainwater and settles on the Earth's surface, where it is included in the formation of sediments and in the accumulation of ice in glaciers and ice-sheets.

The concentration of ^{10}Be at any depth in a sediment or ice-sheet depends on the intensity of the geomagnetic field at the time of deposition. This is because the magnetic field deflects and partially shields the Earth from incident cosmic radiation (Section 11.2.3). Consequently, the ^{10}Be concentration in sediments varies inversely with the field intensity. However, other magnetic fields also affect the cosmic flux. The Sun constantly emits a stream of electrically charged particles, the solar wind, which varies in strength with the activity of the Sun. The solar wind produces magnetic fields that also deflect the incident flux of primary cosmic rays. In this way the solar activity modifies the production rates of cosmogenic isotopes in the atmosphere. The variations in ^{10}Be concentration in a sediment or ice core are therefore the result of changes in both geomagnetic field intensity and solar activity. Measured over geologically short time intervals (e.g., several hundred years) they represent solar activity and can be related to past climatic changes. Measurements of the ^{10}Be concentration over longer time intervals vary inversely with fluctuations in geomagnetic field intensity. As a result, the ^{10}Be concentration in deep-sea sediment cores can be used as a proxy to reconstruct changes in geomagnetic field intensity over hundreds of thousands of years (Fig. 8.11), matching actual observations well.

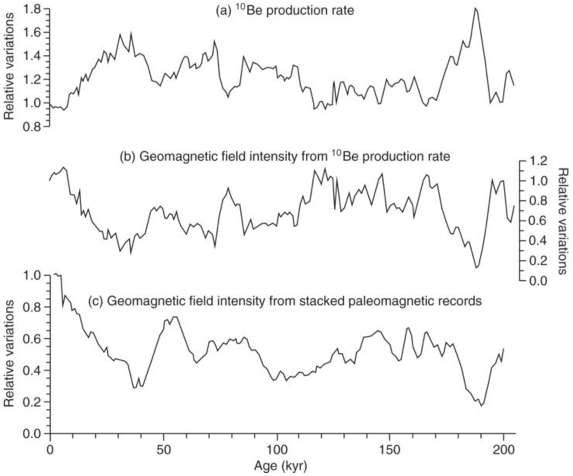


Fig. 8.11 (a) Variations in ^{10}Be production rate estimated from stacked measurements (averaged per kyr) from 19 deep-sea cores; and (b) the proxy record of geomagnetic field variations reconstructed from the ^{10}Be record. Source: Y. Guyodo and J.-P. Valet, Relative variations in geomagnetic intensity from sedimentary records: the past 200,000 years, *Earth Planet. Sci. Lett.*, 143, 23–36, 1996. (c) Stacked profiles of relative variations in geomagnetic field intensity during the same time interval.

Source: Frank, M., et al., A 200 kyr record of cosmogenic radionuclide production rate and geomagnetic field intensity from 10-Be in globally stacked deep-sea sediments. *Earth Planet. Sci. Lett.*, 149, 121–129, 1997.

The flux of cosmic rays that eventually reaches the Earth's surface consists mainly of neutrons and is absorbed in the top one or two meters of exposed rocks, where it produces further cosmogenic nuclides. The two most common of these are ^{26}Al and ^{10}Be, produced from silicon (^{28}Si) and oxygen (^{16}O), respectively, when cosmic rays impact on quartz crystals. The ^{26}Al and ^{10}Be are produced and decay at different rates. By measuring the ratio of their concentrations in a rock sample exposed at the surface, and accounting for factors that affect the cosmic ray flux (e.g., latitude, elevation, exposure to the sky) it is possible to estimate how long the rock sample has been exposed at the surface. After burial below the depth reached by cosmic rays, cosmogenic production ceases and the nuclides decay. The half-life of ^{26}Al is 7.17×10^6 years – half that of ^{10}Be – so the isotopic ratio ^{26}Al/^{10}Be decreases with time. The ratio is measured by means of accelerator mass spectrometry and allows determination of the time of burial. The method is used for dating clastic sediments formed in the late Pliocene and Quaternary (i.e., younger than a few million years).

Vertical ice cores, some penetrating to depths greater than 3 km, have been drilled to basement under the ice-sheets covering Greenland and the Antarctic. Oxygen isotopic ratios (Box 8.1) on the cores have provided important information about past temperatures and greenhouse gas concentrations in the Earth's atmosphere. The dating of these cores is made by a variety of techniques, the simplest of which is to count layers that form due to seasonal variations in the accumulation and compaction of snow. Another dating method is to correlate levels in the core with orbital cyclicities. Radiocarbon dating can measure the ages of the carbon in CO_2 trapped in the ice, but this is useful only for the youngest 50 kyr of a core.

For ice older than this age the recent development of noble gas dating may be important. Among the noble gases, radioactive krypton has two isotopes, ^{85}Kr with a short half-life of 10.76 years and ^{81}Kr with a half-life of 229 kyr. The ^{81}Kr isotope is suitable for determining ages in the 50 kyr to 1.2 Myr range. The ratio of radiokrypton to normal krypton in the atmosphere is only 5.2×10^{-13}, so the technique of atom trap trace analysis (ATTA) developed in atomic physics is essential for age-dating with this isotope. This technique involves capturing atoms with laser beams and inducing them to fluoresce, which occurs when the laser frequency matches the resonant frequency of a chosen atomic transition. The atoms of the isotope isolated in this way are then counted by accelerator mass spectroscopy. Test measurements have been successful in dating ancient ice exposed on the surface of an Antarctic glacier (so-called “blue ice”), whose age was known from stratigraphic matching. The method requires a large mass of ice ($\sim$80 kg/sample) but further refinement may make it applicable to ice cores. The ability to date blue ice with this method already makes it feasible to find and date Earth’s oldest ice, which is expected to have an age of around 1.5 Ma.

8.5 Ages of the Earth and the Solar System

An important source of information concerning the Earth's age is the radiometric dating of meteorites. A *meteor* is a piece of solid matter from space that penetrates the Earth's atmosphere at a hypersonic speed of typically 10–20 km s^{-1}. Atmospheric friction causes it to become incandescent. Outside the Earth's atmosphere it is known as a *meteoroid*; any part that survives passage through the atmosphere and reaches the Earth's surface is called a *meteorite*. Most meteorites are thought to originate in the asteroid belt between the orbits of Mars and Jupiter (see Section 1.5.6), although tracking of entry paths shows that before colliding with Earth they have highly elliptical counterclockwise orbits about the Sun (in the same sense as the planets). Meteorites are often named after the place on Earth where they are found. They can be roughly divided into three main classes according to their composition. Iron meteorites consist of an alloy of iron and nickel; stony meteorites consist of silicate minerals; and iron–stony meteorites are a mixture of the two. The stony meteorites are further subdivided into chondrites and achondrites. *Chondrites* contain small spherules of high-temperature silicates, and constitute the largest fraction (more than 85%) of recovered meteorites. The *achondrites* range in composition from rocks made up essentially of single minerals like olivine to rocks resembling basaltic lava. Each category is further subdivided on the basis of chemical composition. All main types have been dated radiometrically, with most studies being done on the dominant chondrite fraction. There are no obvious age differences between the meteorites of the various groups. Chondrites, achondrites, and iron meteorites consistently yield ages of around 4.45–4.50 Ga (Fig. 8.12).

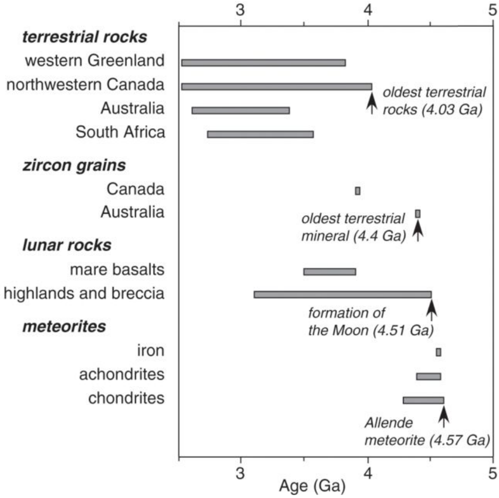


Fig. 8.12 Radiometric age ranges of the oldest terrestrial and lunar rocks and meteorites.

Data source: G.B. Dalrymple, *The Age of the Earth*, 474 pp., Stanford, CA: Stanford University Press, 1991.

In 1969, a large carbonaceous chondritic meteorite entered the Earth's atmosphere and exploded, scattering its fragments over a large area around Pueblito de Allende in Mexico. The Allende meteorite contained refractory inclusions enriched in uranium relative to lead. Refractory inclusions are small aggregates of material that solidify and vaporize at high temperature, and thus can preserve information from the early history of the meteorite, and hence of the solar system. The high U/Pb ratios of the Allende inclusions gave precise $^{207}Pb/^{206}Pb$ dates averaging 4.566 ± 0.002 Ga (Allègre et al., 1995), which is the current best estimate of the age of the solar system.

A further bound on the age of the Earth is placed by the age of the Moon, which is thought to have formed during the accretion of the Earth. Six American manned missions and three Russian unmanned missions obtained samples of the Moon's surface, which have been dated by several isotopic techniques. The ages obtained differ according to the source areas of the rocks. The dark areas of the Moon's surface, the so-called lunar seas or *maria*, were formed when enormous outpourings of basaltic lava filled up low-lying areas – such as the craters formed by large meteoritic impacts – and so created flat surfaces. The lunar volcanism may have persisted until 1 Gyr ago. The light areas on the Moon's surface are rough, extensively cratered highlands that reach elevations of 3000–4000 m above the maria. They represent the top part of the lunar crust and are the oldest regions of the Moon. Frequent collisions with large asteroids early in the Moon's history produced numerous craters and pulverized the lunar crust, leaving impact breccia, rock fragments, and a dust layer a few meters thick, called the lunar *regolith*. Age dates from the highland rocks range from about 3.5 to 4.5 Ga, but the oldest age reported for a lunar rock is 4.51 ± 0.07 Ga, obtained by the Rb–Sr method.

In contrast to the ages of meteorites and the Moon, no rock of comparable age has been preserved on the Earth. Among the oldest terrestrial rocks are the Isua metasediments from western Greenland. They have been extensively studied and yield an age of 3.77 Ga that is consistent with different decay schemes. Even older are the Acasta gneisses in the northwestern part of the Canadian shield, which gave a U–Pb age of 3.96 Ga. Terrestrial rocks from the oldest Precambrian shield areas in Australia, South Africa, South America, and Antarctica give *maximum* ages of 3.4–3.8 Ga.

Older ages are obtained from individual grains of zircon ($ZrSiO_4$), which forms as a trace mineral in granitic magmas. Zircon is a very durable mineral and can survive erosion that destroys the original rock. The zircon grains then become incorporated in sedimentary rocks. Zircon grains from sedimentary rocks in Australia have yielded U–Pb ages of 4.1–4.2 Ga; a fragment of a zircon grain, only 200 μm in diameter, has been dated by the U–Pb method at 4.404 ± 0.008 Ga, making it the oldest dated terrestrial solid. It is also possible to carry out oxygen-isotope analyses on these ancient zircon grains. Although fragmentary, the results from zircon grains are important for understanding conditions on the primeval Earth. The scanty evidence suggests a possible history for the development of the hot early Earth, as follows.

The Sun and its accretionary disk, from which the solar system eventually evolved, formed 4.57 Gyr ago. The accretion of the Earth and formation of its iron core lasted for about the first 100 Myr of Earth's early history. The accretion of Mars, which has only about one-ninth the mass of Earth, would have been completed in the first 30 Myr. Later in Earth's accretion, after perhaps 60 Ma, the Moon was formed as a result of the collision – called the "Giant Impact" – of a Mars-sized planetesimal with the proto-Earth, which was still much smaller than its present size. In the interval up to 4 Gyr ago – called the *Hadean* (for hell-like) era – the early Earth was hot and covered by magma oceans; any crust would melt or be destroyed in the intense meteoritic bombardment. Water, if present, would be vaporized. After about 150 Ma the proto-Earth might have cooled sufficiently so that an early granitic crust, and possibly liquid water, could form. Only a few zircon grains that survived this period remain as possible witnesses. Repeated bombardments may have destroyed the crust and vaporized the water repeatedly in the following 400 Ma. Around 4 Gyr ago the oldest surviving continental crust was formed.

This scenario is speculative and it is not unique. Some oxygen-isotope studies on zircon grains have been interpreted as indicating a cool early Earth. High $\delta^{18}O$ values (see Box 8.1) measured in these zircons imply surface temperatures low enough for liquid water; uniform conditions throughout the Archean era (8.4–2.6 Ga) were deduced. This model of early Earth suggests that meteoritic impacts may have been less intense than hypothesized. The model of a hot early Earth is more popularly accepted, but the true history is not yet known. The events that happened in the first few hundred million years of its history are among Earth's best-kept secrets.

8.6 *Suggestions for Further Reading*

Introductory Level

York, D. 1997. *In Search of Lost Time*. Bristol: Institute of Physics Publishing.

Intermediate and Advanced Level

Dalrymple, G. B. 1991. *The Age of the Earth*. Stanford, CA: Stanford University Press.

Dickin, A. P. 2005. *Radiogenic Isotope Geology* (2nd ed.). Cambridge: Cambridge University Press.

Faure, G. and Mensing, T. M. 2005. *Isotopes: Principles and Applications*. Hoboken, NJ: Wiley.

8.7 Review Questions

1. What is the difference between sidereal time and solar time? Which is longer: a solar day or a sidereal day?

2. Why are sidereal days equal in length but solar days have (slightly) different lengths?

3. Define the following age-dating parameters: (a) decay constant, (b) half-life, (c) isochron.

4. The radioactive carbon method of age-dating is a simple decay analysis. Explain what this statement means. Describe the principle of the method.

5. Describe the principle of a mass spectrometer. What is the Lorentz force?

6. What aspects make the ^{40}K/^{40}Ar method suitable for determining the ages of rocks? What advantages does ^{40}Ar/^{39}Ar dating have over the ^{40}K/^{40}Ar method?

7. What types of materials are suitable for dating with the radioactive carbon method? For what range of ages may it be applied? What are possible problems with the method?

8. Where are the oldest regions of the oceans? Where are the oldest continental regions? Compare the ages of the oldest oceanic and continental regions and account for the difference.

9. Explain the uranium–lead dating method. What is the concordia curve? What is the discordia line? Why is the U–Pb method suitable for dating very old materials, such as Precambrian rocks?

10. Why are zircons important for dating very old rocks? How do the ages of the oldest rocks on Earth compare with the ages of meteorites and the Moon?

8.8 Exercises

1. How many half-lives must elapse before the activity of a radioactive isotope decreases to 1% of its initial value? How long is this time for ^{14}C, which has a decay rate of 1.21×10^{-4} yr^{-1}?

2. Radiocarbon dating of a sample of wood from the tomb of an Egyptian pharaoh gave concentrations of 9.83×10^{-15} mol g^{-1} for ^{14}C and 1.202×10^{-2} mol g^{-1} for ^{12}C. Assuming that the initial ^{14}C/^{12}C ratio in the sample corresponded to the long-term atmospheric ratio of 1.20×10^{-12}, determine the age of the tomb, the percentage of ^{14}C remaining, and the original ^{14}C concentration in the wood.

3. The decay constants of ^{235}U and ^{238}U are $\lambda_{235} = 9.8485 \times 10^{-10}$ yr^{-1} and $\lambda_{238} = 1.55125 \times 10^{-10}$ yr^{-1}. Calculate the half-lives of these uranium isotopes.

4. Assuming that the isotopes ^{235}U and ^{238}U were created in a common event, such as a supernova, and given that their abundances are now in the ratio ^{235}U/^{238}U $= 1/137.88$, calculate how long ago they were created.

5. The analysis of strontium and rubidium isotopes in whole rock samples from a granitic batholith gave the following atomic concentrations in ppm:

Sample	**^{87}Sr**	**^{87}Rb**	**^{86}Sr**
A	2.304	8.831	2.751
B	0.518	29.046	0.420
C	1.619	111.03	1.232
D	1.244	100.60	0.871

(a) Calculate the ^{87}Rb/^{86}Sr and ^{87}Sr/^{86}Sr isotopic ratios for these samples.

(b) Determine the age of the batholith and the initial $^{87}Sr/^{86}Sr$ ratio.

6. Argon–argon dating of muscovite in a Late Cretaceous granite gave the following isotope ratios for the plateau stages during incremental heating:

Maximum heating temperature (°C)	$^{39}Ar/^{36}Ar$	$^{40}Ar/^{36}Ar$
750	1852	8855
830	1790	8439
895	1439	6867
970	3214	15,380
1030	2708	12,970

(a) Calculate the $^{40}Ar/^{39}Ar$ ratios for each incremental heating step.

(b) A calibration constant $J = 0.00964$ was determined for the monitor mineral. Using the $^{40}Ar/^{39}Ar$ ratios from (a) in Eq. (8.18), calculate the apparent ages at each heating step.

(c) Draw an $^{40}Ar/^{39}Ar$ isochron diagram by plotting each $^{40}Ar/^{36}Ar$ ratio as ordinate against the corresponding $^{39}Ar/^{36}Ar$ ratio as abscissa. Draw a best-fitting line – the isochron – through the data points, and determine its slope and intercept.

(d) Compute the age of the muscovite from the slope of the isochron. Is the intercept on the ordinate axis significant?

7. The following isotopic ratios were measured in U–Pb age determinations on three zircon grains extracted from a granite:

Sample	$^{207}Pb/^{235}U$	$^{206}Pb/^{238}U$
Zircon 1	27.4	0.60
Zircon 2	33.3	0.68
Zircon 3	37.9	0.74

(a) With the aid of Eqs. (8.19) and (8.20), plot a concordia diagram on graph paper or with a plotting routine. Enter the measurements from the above table on the graph, and draw the straight discordia line through the points.

(b) Determine the coordinates of the intersection points of the concordia and discordia lines.

(c) Using the coordinates of the upper intersection point together with Eqs. (8.19) and (8.20), calculate the age of formation of the zircons.

(d) Calculate when loss of lead occurred in the zircons.

8. The following isotopic ratios were measured in a K–Ar age determination on an ignimbrite as part of a combined radiometric–paleomagnetic study of geomagnetic polarity.

Sample	**^{40}K/^{36}Ar**	**^{40}Ar/^{36}Ar**
A	4,716,000	822
B	8,069,000	1200
C	12,970,000	1730
D	27,670,000	3280

(a) Plot the isotope ratios, draw the isochron, and compute its slope and intercept.

(b) Calculate the isochron age of the ignimbrite.

(c) Correct the observed ^{40}Ar/^{36}Ar ratios for the initial ^{40}Ar/^{36}Ar concentration, and compute the individual sample ages.

(d) Calculate the mean age and its standard deviation. Compare the mean age with the isochron age.

(e) With reference to the radiometric timescale in Fig. 12.32, what magnetic polarity would you expect the ignimbrite samples to have?

8.9 Computational Exercises

Jupyter notebook computational exercises can be found online for this chapter at www.cambridge.org/FoG3ed (Concordia-discordia diagram).

9

The Earth’s Heat

Preview

The Earth’s internal heat drives geodynamic processes. A primordial part results from cooling of the originally molten planet and part is produced by radioactivity. Temperatures inside the Earth are below the melting point in the crust, mantle, and inner core, but are above the melting point of liquid iron in the outer core. Heat is transported by conduction in the crust, mantle and inner core, and by vigorous convection in the outer core. Deformation of the solid mantle by viscous creep provides a mechanism for mantle convection. We examine the factors that control this process, which provides the power for plate tectonic motions at the surface. The heat flow on continents arises largely from radioactive decay in crustal rocks, but in the oceans results from cooling of lithospheric plates during sea-floor spreading. Plumes of hot mantle material are a possible magma source for volcanic hotspots.

9.1 Introduction

The radiant energy from the Sun, in conjunction with gravitational energy, determines almost all natural processes that occur at or above the Earth's surface. The hot incandescent Sun emits radiation in a very wide range of wavelengths. The radiation incident on the Earth is largely reflected into space; part enters the atmosphere and is reflected by the clouds or is absorbed and re-radiated into space. A very small part reaches the surface, where it is also partly reflected, especially from the water surfaces that cover three-quarters of the globe. Some is absorbed (e.g., by vegetation) and serves as the source of power for various natural cycles. A small fraction is used to heat the Earth's surface, but it only penetrates a short distance – some tens of centimeters in the case of the daily cycle and a few tens of meters for the annual changes. As a result, solar energy has negligible influence on internal terrestrial processes. Systems as diverse as the generation of the geomagnetic field and the motion of global lithospheric plates are ultimately powered by the Earth's internal heat.

The Earth is constantly losing its internal heat. Although small compared to the solar energy that is reflected and re-radiated into space, the loss of internal heat is many times larger than the energy lost by other means, such as the change in Earth's rotation and the energy released in earthquakes (Table 9.1). Tidal friction slows down the Earth's rotation, and the change can be monitored accurately with modern technology such as very long baseline interferometry (VLBI) and the satellite-based global positioning system (GPS). The associated loss of rotational energy can be computed accurately. The elastic energy released in an earthquake can be estimated reliably, and it is known that most of the energy is released in a few large shocks. However, the annual number of large earthquakes is very variable. The number with magnitude $M_s > 7$ varies between about 10 and 40 (see Fig. 7.21), giving estimates of the annual energy release from about 5×10^{17} J to 4×10^{19} J. The energies of tidal deceleration and earthquakes are small fractions of the geothermal flux, which is the most important form of energy originating in the body of the Earth.

Table 9.1 Estimated contributions to the Earth's annual energy budget

Energy source	**Annual energy (joule)**	**Normalized to geothermal flux**
Reflection and re-radiation of solar energy	5.5×10^{24}	≈ 4000
Geothermal flux from the Earth's interior	1.4×10^{21}	1
Rotational deceleration by tidal friction	10^{20}	≈ 0.1
Elastic energy released in earthquakes	3×10^{17}	≈ 0.0002

The Earth's internal heat derives from several sources but for the past 4 Gyr there have been two main sources. One source is primordial heat, the cooling of the Earth since its early history, when internal temperatures were much higher than they are now. The other source is the heat produced by the decay of long-lived radioactive isotopes. This is the main source of the Earth's internal heat, which, in turn, powers all geodynamic processes.

9.2 Thermodynamic Principles

In order to describe thermal energy, it is necessary to define some important thermodynamic parameters. The concepts of temperature and heat are easily – and frequently – confused. Temperature – one of the seven base units in the Systeme Internationale (SI) – is a quantitative measure of the degree of hotness or coldness of an object relative to some standard. Heat is a form of energy that an object possesses by virtue of its temperature. The difference between temperature and heat are illustrated by a simple example. Imagine a container in which the molecules of a gas move around at a certain speed. Each molecule has a kinetic energy proportional to the square of its velocity. There may be differences from one molecule to the next, but it is possible to determine the mean kinetic energy of a molecule. This quantity is proportional to the temperature of the gas. If we add up the kinetic energies of all molecules in the container, we obtain the amount of heat it contains. If heat is added to the container from an external source, the gas molecules speed up, their mean kinetic energy increases, and the temperature of the gas rises.

The change of temperature of a gas is accompanied by changes of pressure and volume. If a solid or liquid is heated, the pressure remains constant but the volume increases. Thermal expansion of a suitable solid or liquid forms the principle of the thermometer for measuring temperature. Although Galileo reputedly invented an early and inaccurate "thermoscope," the first accurate thermometers – and corresponding temperature scales – were developed in the early eighteenth century by Gabriel Fahrenheit (1686–1736), Ferchaut de Réaumur (1683–1757), and Anders Celsius (1701–1744). Their instruments utilized the thermal expansion of liquids and were calibrated at fixed points, such as the melting point of ice and the boiling point of water. The Celsius scale is the most commonly used for general purposes, and it is closely related to the scientific temperature scale.

Temperature apparently has no upper limit. For example, the temperature of the surface of the Sun is less than 10,000 K, but the temperature at its center is around 10,000,000 K and temperatures greater than 100,000,000 K have been achieved in physics experiments. But as heat is removed from an object it becomes more and more difficult to lower its temperature further. The limiting low temperature is often called "absolute zero" and is taken as the zero of the *Kelvin temperature scale*, named in honor of Lord Kelvin. Its divisions are the same as the Celsius scale and the temperature unit is called a *kelvin*. The scale is defined so that the triple point of water – where the solid, liquid, and gaseous phases of water can coexist in equilibrium – is equal to 273.16 kelvins, written 273.16 K.

Heat was imagined by early investigators to be exchanged between bodies by the flow of a mystic fluid, called *caloric*. However, in the mid-nineteenth century James Joule, an English brewer, demonstrated in a series of careful experiments that mechanical energy could be converted into heat. In his famous experiment, falling weights drove a paddle wheel in a container of water, raising its temperature. The increase was tiny, less than 0.3 K, yet Joule was able to compute the amount of energy needed to raise the temperature by 1 K. His estimate of this energy – called the *mechanical equivalent of heat* – was within 5% of our modern value. The unit of energy is called the *joule* in recognition of his pioneering efforts. Originally, however, the unit of heat energy was defined as the amount needed to raise the temperature of one gram of water from 14.5°C to 15.5°C. This unit, the *calorie* (cal), is equivalent to 4.1868 J.

In physics and engineering it is often important to know the change of heat energy in a unit of time, known as the power. The unit of power is the *watt*, named after James Watt, the Scottish engineer who played an important role in harnessing thermal energy as a source of mechanical power. In geothermal problems we are usually concerned with the loss of heat from the Earth per unit area of its surface. This quantity is called the *heat flux* (or more commonly *heat flow*); it is the amount of heat that flows per second across a square meter of surface. The mean heat flow from the Earth is very small and is measured in units of milliwatt per square meter (mW m^{-2}).

The addition of a quantity of heat ΔQ raises the temperature by an amount ΔT, which is proportional to ΔQ. The larger the mass m of the body, the smaller is the temperature change, and a given amount of heat produces different temperature changes in different materials. The amount of heat needed to raise the temperature of 1 kg of a material by 1 K is called its *specific heat*, denoted c_p for a process that occurs at constant pressure (and c_v when it happens at constant volume). These observations are summarized in the equation

$$\Delta Q = c_p m \Delta T. \tag{9.1}$$

The added heat causes a fractional change of volume that is proportional to the temperature change, but which differs from one material to another. The material property is called the *volume coefficient of expansion* α, and is defined by the equation

$$\alpha = \frac{1}{V}\left(\frac{\partial V}{\partial T}\right)_P. \tag{9.2}$$

When thermal energy is added to a system, part is used to increase the internal energy of the system – i.e., the kinetic energy of the molecules – and part is expended as work, for example, by changing the volume. If the change in total energy ΔQ occurs at constant temperature T, we can define a new thermodynamic parameter, the *entropy* S, which changes by an amount ΔS equal to $\Delta Q/T$. Thus we can write

$$\Delta Q = T\Delta S = \Delta U + \Delta W, \tag{9.3}$$

where ΔU is the change of internal energy and ΔW is the work done externally. A thermodynamic process in which heat cannot enter or leave the system is said to be *adiabatic*. The entropy of an adiabatic reaction remains constant: $\Delta S = 0$. This is the case when a process occurs so rapidly that there is no time for heat transfer. An example is the passage of a seismic wave in which the compressions and rarefactions occur too rapidly for heat to be exchanged. The adiabatic temperature gradient in the Earth serves as an important reference for estimates of the actual temperature gradient and for determining how heat is transferred.

9.3 Temperature Inside the Earth

In contrast to the radial distributions of density, seismic velocity, and elastic parameters, which are known with a good measure of reliability, our knowledge of the temperature inside the Earth is still imprecise. The temperature can only be measured in the immediate vicinity of the Earth's surface, in boreholes and deep mines. As early as 1530 Georgius Agricola (the Latinized name of Georg Bauer, a German physician and pioneer in mineralogy and mining) noted that conditions were warmer in deep mines. In fact, near-surface temperatures increase rapidly with depth by roughly 30 K km^{-1}. At this rate, linear extrapolation would give a temperature around 200,000 K at the center of the Earth. This is greater than the temperature of the surface of the Sun and is unrealistically high.

The conditions of high temperature and pressure in the deep interior can be inferred from experiments, and the adiabatic and melting-point temperatures can be computed with reasonable assumptions. Nevertheless, the temperature–depth profile is poorly known and conjectured temperatures have ranged widely. Limits are placed on the actual temperature by the known physical state of the Earth's interior deduced from seismology. The temperature in the solid inner core must be lower than the melting point, while the temperature of the molten outer core is above the melting point. Similarly the temperature in the solid mantle and crust is below the melting point; the asthenosphere has low rigidity because its temperature comes close to the solidus ("softening point"). The relationship of the actual temperature to the melting point determines the rheological behavior of different parts of the Earth's interior (see Section 5.2).

The experimental approach to estimating the variation of temperature with depth combines knowledge obtained from seismology with laboratory results. The travel times of seismic body waves show that changes in mineral structure (phase transitions) occur at certain depths (see Section 7.4.4.1). Important examples are the olivine–spinel transition at 410 km depth and the spinel–perovskite transition at 660 km depth in the upper mantle. The conditions of temperature and pressure (and hence depth) at which these phase transitions take place can be observed in laboratory experiments, so that the temperatures at the transition depths in the Earth can be determined. Similarly, the depth variation of the melting points of mantle rocks and the iron–nickel core can be inferred from laboratory observations at high pressure and temperature. Seismic velocities in the Earth are now so well known that deviations from normal velocities can be determined by seismic tomography (Section 7.4.3.2) and interpreted in terms of temperature anomalies.

9.3.1 The Adiabatic Temperature Gradient

An alternative way of estimating temperature inside the Earth is by using physical equations in which the parameters are known from other sources. In the late nineteenth century, James Clerk Maxwell expressed the laws of thermodynamics in four simple equations involving entropy (*S*), pressure (*p*), temperature (*T*), and volume (*V*). One of these equations is

$$\left(\frac{\partial T}{\partial p}\right)_{\mathrm{S}} = \left(\frac{\partial V}{\partial S}\right)_{\mathrm{P}}. \tag{9.4}$$

The left side is the adiabatic change of temperature with pressure. By combining Eqs. (9.1)–(9.4) and writing density $\rho = m/V$, we get

$$\left(\frac{\partial T}{\partial p}\right)_{\mathrm{S}} = \left(\frac{\partial V}{\partial S}\right)_{\mathrm{P}} = \left(\frac{\partial V}{\partial T}\right)_{\mathrm{P}}\left(\frac{\partial T}{\partial S}\right)_{\mathrm{P}}$$

$$\left(\frac{\partial T}{\partial p}\right)_{\mathrm{S}} = \alpha VT\left(\frac{\partial T}{\partial Q}\right)_{\mathrm{P}} = \frac{\alpha VT}{c_{\mathrm{p}}m} = \frac{\alpha T}{c_{\mathrm{p}}\rho}. \tag{9.5}$$

The variation of temperature with pressure is converted to the variation with depth by substituting $\mathrm{d}p = \rho g\,\mathrm{d}z$. This gives the adiabatic temperature gradient,

$$\left(\frac{\partial T}{\partial z}\right)_{\text{adiabatic}} = T\frac{\alpha g}{c_{\mathrm{p}}}. \tag{9.6}$$

The dependence of density and gravity on depth *z* are known from seismic velocities, and the profiles of *α* and c_{p} can be estimated from laboratory observations (Fig. 9.1). For example, in the lower mantle at a depth of 1500 km, $g = 9.9\ \mathrm{m\ s^{-2}}$, $c_{\mathrm{p}} = 1200\ \mathrm{J\ kg^{-1}\ K^{-1}}$,

$\alpha = 14 \times 10^{-6}$ K^{-1}, and $T = 2400$ K. This gives an adiabatic temperature gradient of about 0.3 K km^{-1}. In the outer core at about 3300 km depth the corresponding values are: $g = 10.1$ m s^{-2}, $c_p = 700$ J kg^{-1} K^{-1}, $\alpha = 14 \times 10^{-6}$ K^{-1}, and $T = 4000$ K, and the adiabatic temperature gradient is about 0.8 K km^{-1}.

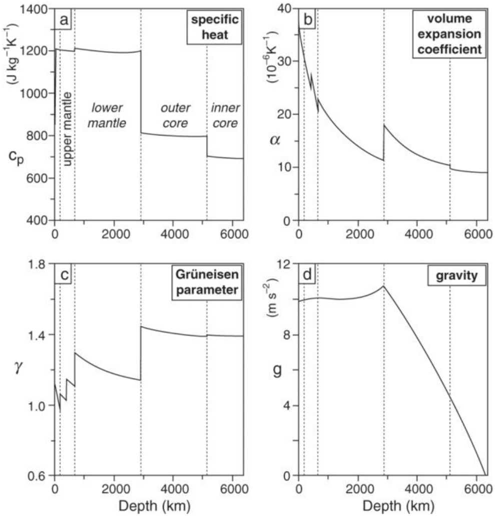


Fig. 9.1 Variations with depth in the Earth of (a) specific heat at constant pressure, (b) volume coefficient of thermal expansion, (c) Grüneisen parameter, and (d) gravity (data source: Stacey and Davis, 2008).

Approximate estimates of adiabatic temperatures inside the Earth can also be obtained with the aid of the Grüneisen thermodynamic parameter, γ. This is a dimensionless parameter, defined as

$$\gamma = \frac{\alpha K_s}{\rho c_p}, \tag{9.7}$$

where K_s is the adiabatic incompressibility or bulk modulus. It is defined in Section 5.1 and Eq. (5.17). By writing dp instead of p and dV/V for the dilatation θ, Eq. (5.17) becomes

$$\mathrm{d}p = -K_s \frac{\mathrm{d}V}{V} = K_s \frac{\mathrm{d}\rho}{\rho}, \tag{9.8}$$

where ρ is the density. Substituting $K_s = \rho\, \partial p/\partial \rho$ in Eq. (9.8) gives

$$\frac{\mathrm{d}T}{\mathrm{d}p} = T\frac{\gamma}{K_s} = T\frac{\gamma \mathrm{d}\rho}{\rho \mathrm{d}p}$$

$$\frac{\mathrm{d}T}{T} = \gamma \frac{\mathrm{d}\rho}{\rho}, \tag{9.9}$$

$$T = T_0 \left(\frac{\rho}{\rho_0}\right)^{\gamma}. \tag{9.10}$$

With this equation, and knowing the temperature T_0 and density ρ_0 at a given depth, the adiabatic temperature can be computed from the density profile in a region where the Grüneisen parameter γ is known. Fortunately, γ is fairly constant within large regions of the Earth's interior (Fig. 9.1). Clearly, Eq. (9.10) cannot be applied across a boundary between these domains, where γ is discontinuous. If T_0 and ρ_0 are known at calibration points, the adiabatic temperature profile may be computed iteratively within a particular depth interval. The temperature profile in the Earth (Fig. 9.2) has steep gradients in the lithosphere, asthenosphere, and in the D″ layer above the core–mantle boundary. It indicates a temperature near 3750 K at the core–mantle boundary and a central temperature of about 6300 K.

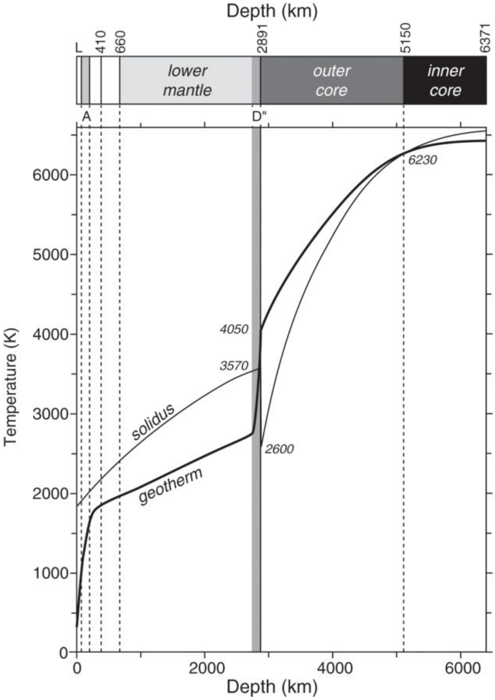


Fig. 9.2 Estimated variations of temperature and melting point with depth (data sources: Stacey and Davis, 2008; Anzellini et al., 2013; and Nomura et al., 2014).

9.3.2 The Melting Point Gradient

Another of Maxwell's thermodynamic equations is

$$\left(\frac{\partial S}{\partial p}\right)_{\mathrm{T}} = -\left(\frac{\partial V}{\partial T}\right)_{\mathrm{P}}. \tag{9.11}$$

This equation can be applied to the effect of pressure on the melting point of a substance (T_{mp}). The heat required to melt a unit mass of the substance is its latent heat of fusion (L), so the change in entropy on the left-hand side of the equation is equal to (mL/T_{mp}). The volume change is the difference between that of the solid phase (V_{S}) and that of the liquid phase (V_{L}), so Eq. (9.11) can be rewritten as

$$\frac{\mathrm{d}T_{\mathrm{mp}}}{\mathrm{d}p} = \frac{T_{\mathrm{mp}}}{mL}\left(V_{\mathrm{S}} - V_{\mathrm{L}}\right). \tag{9.12}$$

This is known to physicists as the *Clausius–Clapeyron equation*. It describes the effect of pressure on the melting point, and it is of interest to us because we can easily convert it to give the variation of melting point with depth, assuming that the pressure is hydrostatic so that $\mathrm{d}p = \rho g\,\mathrm{d}z$ as previously. For a given mass m of the substance we can replace the volumes V_{S} and V_{L} with the corresponding densities ρ_{S} and ρ_{L} of the solid and liquid phases, respectively, so that

$$\frac{1}{T_{\mathrm{mp}}}\frac{\mathrm{d}T_{\mathrm{mp}}}{\mathrm{d}z} = \frac{g}{L}\left(\frac{\rho_{\mathrm{S}}}{\rho_{\mathrm{L}}} - 1\right). \tag{9.13}$$

Again, to obtain the depth distribution of the melting point, the variations of density and gravity with depth in the Earth are needed. At outer core pressures the densities of the solid and liquid phases of iron are about 13,000 kg m^{-3} and 11,000 kg m^{-3}, respectively, and the latent heat of fusion of iron is about 7×10^6 J kg^{-1}, so the gradient of the melting point curve in the outer core is about 1 K km^{-1}, i.e., the melting point in the core increases more steeply with depth than the adiabatic temperature. The computations of the adiabatic and melting temperature curves depend on parameters (e.g., L, α, c_p) that are not known with a great degree of reliability in the Earth, so the temperature profiles (Fig. 9.2) will undoubtedly change and become more secure as basic knowledge improves.

One factor that must still be evaluated is the role of phase transitions in the mantle. The D″ layer just above the core–mantle boundary evidently plays a crucial role in transferring heat from the core to the mantle. It constitutes a thermal boundary layer. Likewise, the lithosphere forms a thermal boundary layer that conveys mantle heat to the Earth's surface. It appears unlikely that the phase transition at 410 km constitutes a thermal boundary layer, but the phase transition at 660 km depth may do so. In the estimated temperature profiles in Fig. 9.2 the phase transitions do not act as thermal boundary layers. Throughout most of the mantle the temperature gradient is assumed to equal the adiabatic gradient, but the mantle is bounded at top and bottom by thermal boundary layers (the lithosphere and D″ layer, respectively) in which the temperature gradient greatly exceeds the adiabatic gradient.

9.4 Heat Transport in the Earth

Heat can be transported by three processes: conduction, convection, and radiation. Conduction and convection require the presence of a material; radiation can pass through space or a vacuum. Conduction is the most significant process of heat transport in solid materials and thus it is very important in the crust and lithosphere. However, it is an inefficient form of heat transport, and when the molecules are free to move, as in a fluid or gas, the process of convection becomes more important. Although the mantle is solid from the standpoint of the rapid passage of seismic waves, the temperature is high enough for the mantle to act as a viscous fluid over long time intervals. Consequently, convection is a more important form of heat transfer than conduction in the mantle. Convection is also the most important form of heat transport in the fluid core, where related changes in the geomagnetic field show that the turnover of core fluid is rapid in geological terms. Radiation is the least important process of heat transport in the Earth. It is only significant in the hottest regions of the core and lower mantle. The absorption of radiant energy by matter increases its temperature and thereby the temperature gradient. Hence, thermal radiation can be taken into account as a modification of the ability of the material to transfer heat by conduction.

9.4.1 Conduction

Thermal conduction takes place by the transfer of kinetic energy between molecules or atoms. A true understanding of the processes involved would require us to invoke quantum theory and the so-called “band theory of solids,” but a general understanding is possible without resorting to such measures. The electrons in an atom that are most loosely bound – the valence electrons – are essentially free of the ionic cores and can move through a material, so transferring kinetic energy. Hence they are called *conduction electrons*. Because electrons are electrically charged, the net movement of conduction electrons also causes an electrical current. Not surprisingly, materials that are good electrical conductors (e.g., silver, copper) also conduct heat well. In this atomic view, the conduction electrons move at very high speeds (~1000 km s^{-1}) but in random directions, so that there is no net energy transfer in any particular direction. In an electrical or temperature field the conduction electrons drift systematically down the slope of the field (i.e., in the direction of the electrical field or temperature gradient). The additional drift velocity is very small (about 0.1 mm s^{-1}), but it passes kinetic energy through the material. This form of conduction is possible in liquids, gases, or solids.

An additional mechanism plays an important role in conduction in solids. The atoms in a solid occupy definite positions that form a lattice with a certain symmetry. The atoms are not stationary, but vibrate at a frequency that is temperature-dependent. The lattice vibrational energy is quantized, forming units called *phonons*. An increase in temperature at one end of a solid raises the lattice vibrational frequency there. Due to the coupling between atoms, the increased vibration is eventually passed through the lattice as an increase in temperature.

The relative importance of electrons and phonons in conducting heat differs from one solid to another. In metals, which contain large numbers of conduction electrons, thermal transport is due largely to the electrons; the lattice conductivity is barely measurable. In an insulator or poor conductor, such as the minerals of the crust and mantle, there are few conduction electrons and thermal conductivity is largely determined by lattice vibrations (phonons).

The transport of heat by conduction in a solid is governed by a simple equation. Consider a solid bar of length L and cross-sectional area A with its ends maintained at temperatures T_1 and T_2, respectively (Fig. 9.3a). Assuming that heat flows only along the bar (i.e., there are no side losses), the net amount of heat (ΔQ) that passes in a given time from the hot end to the cold end depends directly on the temperature difference ($T_2 - T_1$), the area of cross-section (A), and the time of observation (Δt), and inversely on the length of the bar (L). These observations can be summarized in the equation

$$\Delta Q = kA\frac{T_2 - T_1}{L}\Delta t. \tag{9.14}$$

The constant of proportionality, k, is the *thermal conductivity*, which is a property of the material of the bar. If the length of the bar is very small or the temperature change across it is uniform, the ratio $(T_2 - T_1)/L$ is the temperature gradient. We can modify the equation to describe the vertical flow of heat out of the Earth by substituting the vertical temperature gradient, $(\mathrm{d}T/\mathrm{d}z)$, which is also called the *geothermal gradient*. Equation (9.14) can then be rearranged as follows:

$$q_z = -\frac{1}{A}\frac{\mathrm{d}Q}{\mathrm{d}t} = -k\frac{\mathrm{d}T}{\mathrm{d}z}. \tag{9.15}$$

In this equation q_z is the *heat flux*, defined as the flow of heat per unit area per second. The negative sign is needed to account for the direction of the heat flow; if temperature increases in the downward direction of the z-axis, the flow of heat from high to low temperature is upward. The mean value of the undisturbed geothermal gradient near to the Earth's surface is about 30°C km^{-1}, with low values of around 10°C km^{-1} in ancient crust and high values of around 50°C km^{-1} in young active zones.

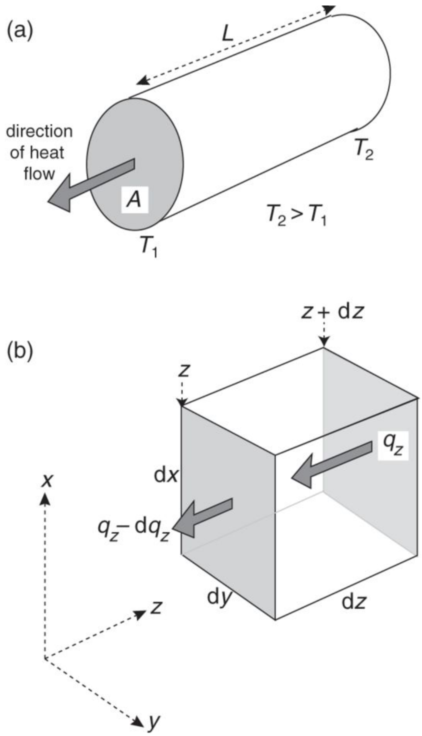


Fig. 9.3 (a) Conduction of heat *Q* through a bar of length *L* and cross-sectional area *A*, with its ends kept at temperatures T_1 and T_2 ($> T_1$). (b) Heat flux entering (q_z) and leaving ($q_z - \mathrm{d}q_z$) a short bar of length d*z*.

The change of temperature within a body is described by the heat conduction equation, which is solved in Section 9.6 for special situations that are of interest for the transfer of thermal energy in the Earth. Conduction is a slow and less effective means of heat transport than convection. It is important in the rigid crust and lithosphere, where convection cannot take place. However, it cannot be neglected in the fluid core, which is metallic and therefore a good conductor. A significant part of the core's heat is conducted out of the core along the adiabatic temperature gradient. The remainder, in excess of the conductive heat flow, is transported by convection currents.

9.4.2 Convection

Suppose that a small parcel of material at a certain depth in the Earth is in thermal equilibrium with its surroundings. If the parcel is displaced vertically upward without gaining or losing heat, it experiences a drop in pressure accompanied by a corresponding loss of temperature. If the new temperature of the parcel is the same as that of its surroundings at the new depth, the conditions at each depth are in adiabatic equilibrium. The variation of temperature with depth then defines the adiabatic temperature curve.

Now suppose that the real temperature increases with depth more rapidly than the adiabatic temperature gradient. The temperature loss of the upwardly displaced parcel is due to the change in pressure, which will be the same as in the previous case. But the real temperature has dropped by a larger amount, so the parcel is now hotter and therefore less dense than its surroundings. Its buoyancy causes it to continue to rise until it reaches a level where it is in equilibrium or can rise no further. Meanwhile, the volume vacated by the displaced parcel is occupied by adjacent material. Conversely, if a parcel of material is displaced downward, it experiences adiabatic increases in pressure and temperature. The temperature increase is less than required by the real temperature gradient, so the parcel remains cooler than its surroundings and sinks further. A pattern of cyclical behavior arises in which material is heated up and rises, while cooler material sinks to take its place, and is in turn heated up and rises, and so on. The process is called *thermal convection* and the physical transportation of material and heat is called a *convection current*.

The difference between the real and adiabatic temperature gradients is the *superadiabatic* temperature gradient, θ. For thermal convection to take place in a fluid, θ must be positive. Suppose that the temperature at a certain depth exceeds the adiabatic temperature by an amount ΔT. The temperature excess causes a volume V of fluid to expand by an amount proportional to the volume coefficient of expansion, α; this causes a mass deficiency of ($V\rho\alpha\,\Delta T$). Archimedes' principle applies, so the hot volume V experiences a buoyancy force given by:

$$F_B = V\rho g\alpha\,\Delta T. \tag{9.16}$$

Two effects inhibit the hot volume from rising. First, some of the heat that would contribute to the buoyancy is removed by thermal conduction; the efficacy of this process is expressed by the *thermal diffusivity* κ of the material, which depends on its density ρ, thermal conductivity k, and specific heat c_p (see Section 9.6). Second, as soon as the overheated volume of fluid begins to rise, it experiences a resisting drag due to the *viscosity* η of the fluid. The effects combine to produce a force, proportional to $\kappa\eta$, which opposes convection. If the volume V involved in the convection has a typical dimension D, so that $V \sim D^3$, we can define a dimensionless number Ra, the Rayleigh number, which is proportional to the ratio of the buoyancy force to the diffusive–viscous force:

$$\mathrm{Ra} = \frac{g\rho\alpha\,\Delta T}{k\eta} D^3. \tag{9.17}$$

Initially, heat passes through the material by conduction, but the diffusion takes some time. If the heat flux is large enough, it cannot diffuse entirely. The temperature rises above the adiabatic and buoyancy forces develop. For convection to occur, the buoyancy forces must dominate the resisting forces. This does not happen until the Rayleigh number exceeds a critical value, which is determined additionally by the boundary conditions and the geometry of the convection. For example, the condition for the onset of convection in a thin horizontal fluid layer, heated from below and with the top and bottom surfaces free from stress, was shown by Lord Rayleigh in 1916 to depend on the value of Ra given by

$$\mathrm{Ra} = \frac{g\alpha\,\theta}{k\nu} D^4 . \tag{9.18}$$

The superadiabatic temperature difference ΔT in Eq. (9.17) – the amount by which the temperature difference exceeds the adiabatic temperature difference – has been replaced here by the product (θD), where θ is the superadiabatic temperature gradient $(\theta = \Delta T/D)$, D is the layer thickness, and ν (equal to η/ρ) is the kinematic viscosity. Rayleigh's theory showed that convection begins in the flat layer if Ra is greater than $(27\pi^4/4) = 658$. In cases with different boundary conditions, or for convection to occur in a spherical shell, the critical Rayleigh number is higher. However, convection generally initiates if Ra is of the order of 10^3 and when Ra reaches around 10^5 heat transport is almost entirely by convection, with little being transferred by diffusion.

For convection to occur, the real temperature gradient must exceed the adiabatic gradient. However, the loss of heat by convection reduces the difference between the gradients. Accordingly, the adiabatic gradient evolves as a convecting fluid cools. An important effect of convection is to keep the temperature gradient close to the adiabatic gradient. This condition is realized in the Earth's fluid core, where convection is the major mechanism of heat transport. Thermal convection is augmented by *compositional convection* related to the solidification of the inner core. The core fluid is made up of iron, nickel, and lower-density elements, e.g., sulfur. Solidification of the inner core separates the dense iron from the lower-density elements at the inner core boundary. Being less dense than the core fluid, the residual materials experience an upward buoyancy force, resulting in a cycle of compositionally driven convection. Thermal and compositional convection in the Earth's core each act as a source of the energy needed to drive the geomagnetic field.

Convection is the most important process of thermal transport in the fluid core, but it is also important in the mantle. The material of the Earth's mantle is rigid to the short-lived passage of seismic waves but is believed to yield slowly over long periods of time (Section 5.3.3). Although the mantle viscosity is high, the timescale of geological processes is so long that long-term flow can take place. The flow patterns are dominated by thermal convection and are influenced by the presence of thermal boundary layers, which the flowing material cannot cross. However, convection is a more effective mechanism than conduction and it is thought to be the dominant process of heat transfer in the mantle (Section 9.9).

A further process of heat transfer that involves bodily transport of matter is *advection*. This can be regarded as a form of forced convection. Instead of being conveyed by thermally produced buoyancy, advected heat is transported in a medium that is itself driven by other forces. For example, in a thermal spring the flow of water is due to hydraulic forces and not to density differences in the hot water. Similarly, volcanic eruptions transport advected heat along with the lava flow, but this is propelled by pressure differences rather than by buoyancy.

9.4.3 Radiation

Atoms can exist in many distinct energy states. The most stable is the ground state, in which the energy is lowest. When an atom changes from an excited state to a lower-energy state, it is said to undergo a transition. Energy corresponding to the difference in energy between the states is emitted as an electromagnetic wave, which we call radiation. Quantum physics teaches that the radiant energy emitted consists of a discrete number of fundamental units, called *quanta*. The particular wavelength of the electromagnetic radiation associated with a transition is proportional to the energy difference between the two states. If several different transitions are taking place simultaneously, the body emits a spectrum of wavelengths. Radio signals, heat, light, and X-rays are examples of electromagnetic radiation that have different wavelengths. The electromagnetic wave consists of fluctuating electric and magnetic fields, which need no medium for their passage. For this reason, radiation can travel through space or a vacuum. In materials it may be scattered or absorbed, depending on its wavelength. Heat radiation corresponds to the *infrared* part of the electromagnetic spectrum with wavelengths just longer than those of visible light.

The radiation of a commonplace hot object depends on factors that are difficult to assess. Classical physics fails to explain adequately the absorption and emission of radiation. To provide an explanation, physicists introduced the concept of a *black body* as a perfect absorber and emitter of radiation. At any temperature it emits a continuous spectrum of radiation; the frequency content of the spectrum does not depend on the material composition of the body but only on its temperature. An ideal black body does not exist in practice, but it can be approximated by a hollow container that has a small hole in its wall. When the container is heated, the radiation escaping through the hole – so-called cavity radiation – is effectively black-body radiation. In 1879 Josef Stefan pointed out that the loss of heat by radiation from a hot object is proportional to the fourth power of the absolute temperature. If R represents the radiant energy per second emitted per unit area of the surface of the body at temperature T, then

$$R = \sigma T^4, \tag{9.19}$$

where σ, known as Stefan's constant or the Stefan–Boltzmann constant, has the value 5.670367×10^{-8} W m^{-2} K^{-4}.

In 1900, Max Planck, professor of physics at the University of Berlin, proposed that, in contrast to classical physics, an oscillator could only have discrete amounts of energy. This was the birth of quantum theory. The energy of an oscillator of frequency ν is equal to the product $h\nu$, where the universal constant h (known as Planck's constant) has the value 6.626070×10^{-34} J s. The application of quantum principles to black-body radiation provides a satisfactory explanation of Stefan's law and allows Stefan's constant to be expressed in terms of other fundamental physical constants.

Radiation is reflected and refracted in a transparent medium wherever the refractive index n changes; energy is transferred to the medium in each of these interactions. The transparency of the medium is determined by the opacity e, which describes the degree of absorption of electromagnetic radiation. The opacity is wavelength-dependent. In an ionic crystal the absorption of infrared radiation is large. It alters the vibrational frequency, and thereby influences the ability of the crystal lattice to transport heat by conduction. Thus the effect can be taken into account by increasing the conductivity by an extra radiative amount, k_r, given by

$$k_r = \frac{16}{3} \frac{n^2 \sigma}{e} T^3. \tag{9.20}$$

The T^3-dependence in this expression suggests that radiation might be more important than lattice conductivity in the hotter regions of the Earth. In fact, other arguments lead to the conclusion that this is probably not the case in the upper mantle, because the effect of increasing temperature is partly offset by an increase in the opacity, e. The lower mantle is believed to have a high density of free electrons, which efficiently absorb radiation and raise the opacity. This may greatly reduce the efficacy of heat transfer by radiation in the mantle.

9.5 Sources of Heat in the Earth

The interior of the Earth is losing heat via geothermal flux at a rate of about 4.4×10^{13} W, which amounts to 1.4×10^{21} J yr^{-1} (see Table 9.1). The heat is brought to the surface in different ways. The creation of new lithosphere at oceanic ridges releases the largest fraction of the thermal energy. A similar mechanism, the spreading of the sea floor, releases heat in the marginal basins behind island arcs. Rising plumes of magma originating deep in the mantle bring heat to the surface, where they break through the oceanic or continental lithosphere at "hotspots," characterized by intense localized volcanic activity. These important thermal fluxes are superposed on a background consisting of heat flowing into and through the lithosphere from deeper parts of the Earth. There are two main sources of the internal heat. Part of it is probably due to the slow cooling of the Earth from an earlier hotter state; part is generated by the decay of long-lived radioactive isotopes.

The early thermal history of the Earth is obscure and a matter of some speculation. According to the cold accretion model of the formation of the planets (see Section 1.4), colliding bodies in a primordial cloud of dust and gas coalesced by self-gravitation. The gravitational collapse released energy that heated up the Earth. When the temperature reached the melting point of iron, a liquid core formed, incorporating also nickel and possibly sulfur or another light element associated with iron. The differentiation of a denser core and lighter mantle from an initially homogeneous fluid must have released further gravitational energy in the form of heat. The dissipation of Earth's initial heat still has an important effect on internal temperatures.

Energy released by short-lived radioactive isotopes may have contributed to the initial heating, but the short-lived isotopes would be consumed quite early. The heat generated by long-lived radioactive isotopes has been an important heat source during most of Earth's history. These isotopes separated into two fractions: some, associated with heavy elements, sank into the core; some, associated with lighter elements, accumulated in the crust. The present distribution of radiogenic sources within the differentiated Earth is uneven. The highest concentrations are in the rocks and minerals of the Earth’s crust, while the concentrations in mantle and core materials are low. However, continuing generation of heat by radioactivity in the deep interior, though small, may influence internal temperatures.

9.5.1 Radioactive Heat Production

When a radioactive isotope decays, it emits energetic particles and γ-rays. The two particles that are important in radioactive heat production are α-particles and β-particles. The α-particles are equivalent to helium nuclei and are positively charged, while β-particles are electrons. In order to be a significant source of heat, a radioactive isotope must have a half-life comparable to the age of the Earth, the energy of its decay must be fully converted to heat, and the isotope must be sufficiently abundant. The main isotopes that fulfill these conditions are ^{238}U, ^{235}U, ^{232}Th, and ^{40}K. The isotope ^{235}U has a shorter half-life than ^{238}U (see Table 8.1) and releases more energy in its decay. In natural uranium the proportion of ^{238}U is 99.28%, that of ^{235}U is about 0.71%, and the rest is ^{234}U. The abundance of the radioactive isotope ^{40}K in natural potassium is only 0.01167%, but potassium is a very common element and its heat production is not negligible. The amounts of heat generated per second by these elements (in μW kg$_{-}{}^{1}$) are: natural uranium, 95.2; thorium, 25.6; and natural potassium, 0.00348. The heat Q_r produced by radioactivity in a rock that has concentrations C_U, C_{Th}, and C_K, respectively, of these elements is

$$Q_r = 95.2C_U + 25.6C_{Th} + 0.00348C_K.$$

(9.21)

Rates of radioactive heat production computed with this equation are shown for some important rock types in Table 9.2. Chondritic meteorites, made up of silicate minerals like olivine and pyroxene, are often taken as a proxy for the initial composition of the mantle; likewise, the olivine-dominated rock dunite represents the ultramafic rocks of the upper mantle. It is apparent that very little heat is produced by radioactivity in the mantle or in the basaltic rocks that dominate the oceanic crust and lower continental crust. The greatest concentration of radiogenic heat sources is in the granitic rocks in the upper continental crust. Multiplying the radioactive heat production values in the last column of Table 9.2 by the rock density gives the radiogenic heat generated in a cubic meter of the rock, *A*. If we assume that all the heat generated in a rock layer of thickness *D* meters escapes vertically, the amount crossing a square meter at the surface per second (i.e., the radioactive component of the heat flow) is *D*A. For example, a 1 km thick layer of granite contributes about 3 mW m^{-2} to the continental heat flow. The figures suggest that the 10–20 km thick upper crust produces about one-half of the mean continental heat flow, which is 65 mW m^{-2}.

Table 9.2 Estimates of radioactive heat production in selected rock types, based on heat production rates (Rybach, 1976, 1988) and isotopic concentrations (Stacey, 1992)

Rock type	**Concentration (ppm by weight)**			**Heat production (10^{-11}W kg^{-1})**			
	U	**Th**	**K**	**U**	**Th**	**K**	**Total**
Granite	4.6	18	33,000	43.8	46.1	11.5	101
Alkali basalt	0.75	2.5	12,000	7.1	6.4	4.2	18
Tholeiitic basalt	0.11	0.4	1500	1.05	1.02	0.52	2.6
Peridotite, dunite	0.006	0.02	100	0.057	0.051	0.035	0.14
Chondrites	0.015	0.045	900	0.143	0.115	0.313	0.57
Continental crust	1.2	4.5	15,500	11.4	11.5	5.4	28
Mantle	0.025	0.087	70	0.238	0.223	0.024	0.49

In fact the relative importance of radiogenic heat in the crust is variable from one region to another. A region in which the heat flow is linearly related to the heat produced by radioactivity is called a *heat-flow province*. Some examples of heat-flow provinces are western Australia, the Superior Province in the Canadian shield, and the Basin-

and-Range Province in the western United States. As shown in Fig. 9.4 each province is characterized by a different linear relation between q and A, such that

$$q = q_r + DA. \tag{9.22}$$

The parameters q_r and D typify the heat-flow province. The intercept of the straight line with the heat-flow axis, q_r, is called the *reduced heat flow*. This is the heat flow that would be observed in the province if there were no radiogenic crustal heat sources. It is due in part to the heat flowing from deeper regions of the Earth into the base of the crustal layer, and partly to cooling of the originally hotter upper crustal layer. Investigations in different heat-flow provinces show that the reduced heat flow averages about 55% of the mean measured heat flow in a province (Fig. 9.5).

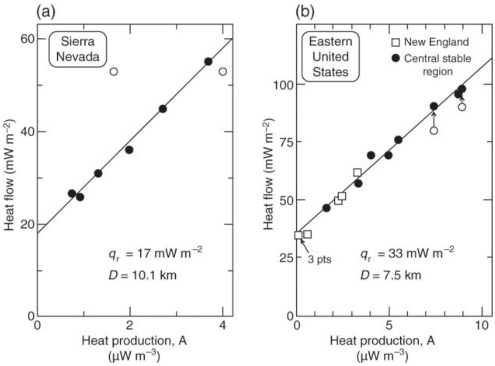


Fig. 9.4 Dependence of surface heat flux on radioactive heat generation in two heat-flow provinces: (a) Sierra Nevada, (b) eastern United States.

Data source: R.F. Roy, D.D. Blackwell and F. Birch, Heat generation of plutonic rocks and continental heat flow provinces, *Earth Planet. Sci. Lett.* 5, 1–12, 1968.

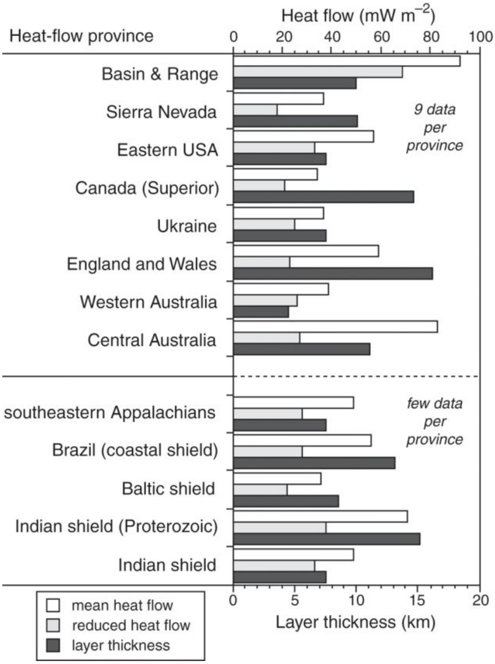


Fig. 9.5 Mean heat flow, reduced heat flow, and characteristic thickness of the layer of radioactive heat production in several heat-flow provinces.

Data source: I. Vitorello and H.N. Pollack, On the variation of continental heat flow with age and the thermal evolution of continents, *J. Geophys. Res.* 85, 983–995, 1980.

The simplest interpretation of D is to regard it as a characteristic thickness of crust involved in radioactive heat production. This assumes that the radiogenic heat sources are distributed uniformly in a crustal slab of constant thickness, which is an unlikely situation. A more likely model is that the radioactive heat generation decreases with depth. Assuming an exponential decrease, the heat production $A(z)$ at depth z is related to the surface heat generation A_0 as

$$A(z) = A_0 e^{-z/D}, \tag{9.23}$$

where D is a characteristic depth (the depth at which $A(z)$ has decreased to e^{-1} of its surface value). Integrating from the surface to infinite depth gives the total radioactive heat production:

$$\int_0^\infty A(z)\mathrm{d}z = A_0\int_0^\infty e^{-z/D}\mathrm{d}z = DA_0, \tag{9.24}$$

which is the same as for the uniform distribution. The infinite depth limit to the exponential distribution is obviously unrealistic. If the radiogenic sources are distributed in a layer of finite thickness *s*, the integration becomes

$$A_0\int_0^s e^{-z/D}\mathrm{d}z = DA_0(1 - e^{-s/D}). \tag{9.25}$$

If *s* is greater than three times *D*, this expression differs from DA_0 by less than 5%. The value of *D* estimated from studies of heat flow and radioactive heat generation averages about 10 km, but varies from ~4 km to ~16 km (Fig. 9.5).

The three main sources of the Earth's surface heat flow are (1) heat flowing into the base of the lithosphere from the deeper mantle; (2) heat lost by cooling of the lithosphere with time; and (3) radiogenic heat production in the crust. The contributions are unequal and different in the oceans and continents (Table 9.3). The most obvious disparity is in the relative importance of lithospheric cooling and radioactivity. The lithosphere is hot when created at oceanic ridges and cools slowly as it ages. The loss of heat by lithospheric cooling is most pronounced in the oceanic crust, which moreover contains few radiogenic heat sources. In contrast, the older continental lithosphere has lost much of the early heat of formation, and the higher concentration of radioactive minerals increases the importance of radiogenic heat production. Regardless of its source, the passage of heat through the rigid outer layers takes place predominantly by conduction, although in special circumstances, such as the flow of magma in the crust or hydrothermal circulation near to oceanic ridges, convection also plays an important role.

Table 9.3 Approximate relative contributions (%) of the main sources of heat flow in oceanic and continental lithosphere

Heat source	**Contribution to heat flow in:**	
	Continents (%)	**Oceans (%)**
Cooling of the lithosphere	20	85
Heat flow from below the lithosphere	25	10
Radiogenic heat:	55	5
Upper crust	40	–
Rest of lithosphere	15	–

9.6 The Heat Conduction Equation

Jean Baptiste Joseph Fourier (1768–1830), a noted French mathematician and physicist, published the theory of heat conduction in 1822. Here we consider the example of one-dimensional heat flow, which typifies many interesting problems involving the flow of heat in a single direction. The equation of heat conduction is most easily developed for this case, from which it can be readily extended to three-dimensional heat flow.

Consider the flow of heat in the negative direction of the z-axis through a small rectangular prism with sides dx, dy, and dz (see Fig. 9.3b). We will assume that there are no sources of heat inside the box. Let the amount of heat entering the prism at (z + dz) be Q_z. This is equal to the heat flow q_z multiplied by the area of the surface it flows across (dx dy) and by the duration of the flow (dt). The heat leaving the box at z is $Q_z - \mathrm{d}Q_z$, which can be written $Q_z - (\mathrm{d}Q_z/\mathrm{d}z)\,\mathrm{d}z$. The increase in heat in the small box is the difference between these amounts:

$$\frac{\mathrm{d}Q_z}{\mathrm{d}z}\mathrm{d}z = \frac{\mathrm{d}q_z}{\mathrm{d}z}\mathrm{d}z(\mathrm{d}x\ \mathrm{d}y)\mathrm{d}t = k\frac{\mathrm{d}^2T}{\mathrm{d}z^2}\mathrm{d}V\ \mathrm{d}t, \tag{9.26}$$

where dV is the volume of the box (dx dy dz). Note that Q_z, q_z, and T are all understood to decrease in the direction of flow, so we have substituted for q_z from Eq. (9.39) without using the negative sign. The heat increase in the box causes its temperature to rise by an amount (dT), determined by the specific heat at constant pressure (c_p) and the mass of material (m) in the box. Using Eq. (9.25), we write

$$c_\mathrm{p}m\ \mathrm{d}T = c_\mathrm{p}\rho\ \mathrm{d}V\ \mathrm{d}T, \tag{9.27}$$

where ρ is the density of the material in the box. If we equate Eq. (9.26) and Eq. (9.27) with the amount of heat left in the box, we get the equation of heat conduction

$$\frac{dT}{dt} = \frac{k}{\rho c_p}\frac{d^2T}{dz^2}, \tag{9.28}$$

or, changing to partial differentials,

$$\frac{\partial T}{\partial t} = \kappa\frac{\partial^2 T}{\partial z^2}, \tag{9.29}$$

where κ (= $k/(\rho c_p)$) is called the *thermal diffusivity*; it has the dimensions $m^2\ s^{-1}$. The equation is written with partial differentials because the temperature T is a function of both time and position: $T = T(z, t)$. This just means that, on the one hand, the temperature at a certain position changes with time, and, on the other hand, the temperature at any given time varies with position in the body.

The same arguments can be applied to the components of heat flow through the box in the x- and y-directions. We obtain the three-dimensional heat-flow equation (also called the *diffusion equation*):

$$\frac{\partial T}{\partial t} = \kappa\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2}\right). \tag{9.30}$$

The equation may be solved for any set of boundary conditions by using the method of *separation of variables* (Box 9.1). Two important situations involving the flow of heat across the Earth's surface are the heating of Earth's surface by solar energy and the cooling of hot lithosphere. These can be handled to a first approximation as problems of one-dimensional heat conduction.

Box 9.1 Method of Separation of Variables

The method of separation of variables is encountered in the solution of several geophysical problems. It may be illustrated by the case of one-dimensional heat flow, described by Eq. (9.29). The coordinates in this equation are independent, i.e., each can take any value regardless of the other. Assume that the temperature $T(z, t)$ can be written as the product of $Z(z)$, a function of position only, and $\theta(t)$, a function of time only:

$$T(z,t) = Z(z)\theta(t). \tag{1}$$

The actual solution for $T(z, t)$ usually does not have this final form, but once $Z(z)$ and $\theta(t)$ are known they can be combined to fit the boundary conditions. On substituting in Eq. (9.29) we get

$$Z\frac{\partial\theta}{\partial t} = \kappa\frac{\partial^2 Z}{\partial z^2}\theta. \tag{2}$$

Dividing both sides by $Z\theta$ gives

$$\frac{1}{\theta}\frac{\partial\theta}{\partial t} = \kappa\frac{1}{Z}\frac{\partial^2 Z}{\partial z^2}. \tag{3}$$

Note that in Eq. (3) the left side depends only on t and the right side depends only on *z*, and that *z* and *t* are independent variables. On substituting a particular value for the time *t* the left side becomes a numerical constant. We have not restricted *z*, which varies independently. Eq. (3) requires that the variable right side remains equal to the numerical constant for any value of *z*. Conversely, if we substitute a particular value for *z*, the right side becomes a (new) numerical constant and the variable left side must equal this constant for any *t*. The identity inherent in Eq. (3) implies that each side must be equal to the same *separation constant*. Let this constant be C. Then,

$$\frac{1}{\theta}\frac{\partial \theta}{\partial t} = C$$

$$\kappa \frac{1}{Z}\frac{\partial^2 Z}{\partial z^2} = C. \quad (4)$$

The value of the separation constant in a particular problem is determined by the boundary conditions of the problem.

9.6.1 Penetration of External Heat into the Earth

Solar energy is by far the largest of the Earth’s energy sources (Table 9.1). In order to determine the geothermal flux from the Earth’s interior, it is important to understand the effects of solar energy that reaches the Earth’s surface. Rocks at the Earth’s surface heat up during the day and cool down at night. The effect is not restricted to the immediate surface, but affects a volume of rock near the surface. Similarly, the mean surface temperature varies throughout the year with the changing seasons. The heat conduction equation allows us to estimate what depths are affected by these cyclical temperature variations.

Box 9.2 One-Dimensional Heat Conduction

The surface-temperature variation $T_0 \cos \omega t$ can be expressed with the aid of complex numbers (Box 4.6) as the real part of $T_0\, e^{i\omega t}$. Let z represent the depth of a point below the surface. The heat conduction equation, Eq. (9.29), can be separated and written as two parts equal to the same constant. We want to match boundary conditions with the surface disturbance $T_0\, e^{i\omega t}$, so we write the separation constant as $i\omega$. Applying the method of separation of variables (Box 9.1), we get for the two parts of the solution

$$\frac{1}{\theta}\frac{\partial \theta}{\partial t} = i\omega, \tag{1}$$

$$\kappa \frac{1}{Z} \frac{\partial^2 Z}{\partial z^2} = \mathrm{i}\omega .$$

(2)

The solution of Eq. (1) gives the time-dependence of the temperature as

$$\theta = \theta_0 \mathrm{e}^{\mathrm{i}\omega t} .$$

(3)

In order to find the depth-dependence, we can rewrite Eq. (2) as

$$\frac{\partial^2 Z}{\partial z^2} - \mathrm{i}\left(\frac{\omega}{\kappa}\right) Z = 0.$$

(4)

Writing $-n^2 = \mathrm{i}\frac{\omega}{\kappa}$, this is equivalent to the harmonic equation

$$\frac{\partial^2 Z}{\partial z^2} + n^2 Z = 0,$$

(5)

which has the solutions

$$Z = Z_0 \mathrm{e}^{inz} \qquad \text{and} \qquad Z = Z_1 \mathrm{e}^{-inz},$$

(6)

where

$$\mathrm{i}n = \sqrt{\mathrm{i}\frac{\omega}{\kappa}} = \sqrt{\frac{\omega}{2\kappa}}(1+\mathrm{i}). \tag{7}$$

We get two possible solutions for the depth variation:

$$Z = Z_1 \mathrm{e}^{\mathrm{i}nz} = Z_1 \mathrm{e}^{\sqrt{\frac{\omega}{2\kappa}}(1+\mathrm{i})z}$$
$$Z = Z_0 \mathrm{e}^{-\mathrm{i}nz} = Z_0 \mathrm{e}^{-\sqrt{\frac{\omega}{2\kappa}}(1+\mathrm{i})z}. \tag{8}$$

The temperature must decrease with increasing depth *z* below the heated surface, so only the second solution is acceptable. Combining the solutions for *θ* and *Z*, we get

$$T(z,t) = Z_0 \mathrm{e}^{-\sqrt{\frac{\omega}{2\kappa}}(1+\mathrm{i})z} \theta_0 \mathrm{e}^{\mathrm{i}\omega t}, \tag{9}$$

$$T(z,t) = Z_0 \theta_0 \mathrm{e}^{-\sqrt{\frac{\omega}{2\kappa}}z} \mathrm{e}^{\mathrm{i}\left(\omega t - \sqrt{\frac{\omega}{2\kappa}}z\right)}. \tag{10}$$

The surface-temperature variation $T_0 \cos \omega t$ was expressed as the real part of $T_0\ \mathrm{e}^{\mathrm{i}}\omega t$. Taking the real part of Eq. (10), we get for the temperature $T(z,t)$ at time t and depth z

$$T(z,t) = Z_0\theta_0 e^{-\sqrt{\frac{\omega}{2\kappa}}z} \cos\left(\omega t - \sqrt{\frac{\omega}{2\kappa}}z\right).$$

(11)

Writing $T_0 = Z_0\theta_0$ and $d = \sqrt{\frac{2\kappa}{\omega}}$, the equation reduces to

$$T(z,t) = T_0 e^{-\frac{z}{d}} \cos\left(\omega t - \frac{z}{d}\right).$$

(12)

The parameter d is called the *decay depth* of the temperature. At this depth the amplitude of the temperature fluctuation is attenuated to 1/e of its value on the surface. Equation (12) can also be written in the form

$$T(z,t) = T_0 e^{-\frac{z}{d}} \cos\omega(t - t_d),$$

(13)

where the phase difference, or *delay time*,

$$t_d = \frac{z}{\omega d},$$

(14)

represents the length of time by which the temperature at depth z lags behind the surface temperature.

Suppose that the surface temperature of the Earth varies cyclically with angular frequency ω, so that at time t it is equal to $T_0 \cos \omega t$, where T_0 is the peak temperature during a cycle. The temperature at time t and depth z is obtained by solving the one-dimensional heat equation (Box 9.2) and is given by

$$T(z,t) = T_0 e^{-z/d} \cos\left(\omega t - \frac{z}{d}\right),$$

(9.31)

where $d = (2\kappa/\omega)^{1/2}$ is the *decay depth* of the temperature. At a depth of $5d$ the amplitude is less than 1% of the surface value and is effectively zero. Note that d depends inversely on the frequency, so long-period fluctuations penetrate more deeply than rapid fluctuations. This is illustrated by a quick comparison of the decay depths for daily and annual temperature variations in the same ground,

$$\frac{d_{\text{annual}}}{d_{\text{daily}}} = \sqrt{\frac{\omega_{\text{daily}}}{\omega_{\text{annual}}}} = \sqrt{365} = 19.1,$$

(9.32)

i.e., the annual variation penetrates about 19 times the depth of the daily variation (Fig. 9.6). Moreover, the temperature at any depth varies with the same frequency ω as the surface variation, but it experiences a phase shift or delay, reaching its maximum value progressively later than the surface effect (Fig. 9.6).

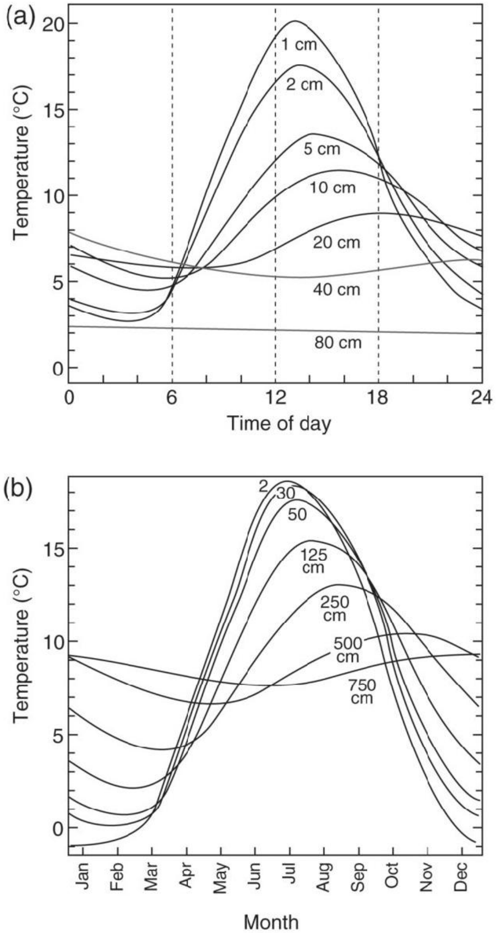


Fig. 9.6 Temperature variations at various depths in a sandy soil: (a) daily fluctuations, (b) annual (seasonal) variations.

As representative values for crustal rocks we take: density, $\rho = 2650\ \text{kg m}^{-3}$; thermal conductivity, $k = 2.5\ \text{W m}^{-1}\ \text{K}^{-1}$; and specific heat $c_p = 700\ \text{J kg}^{-1}\ \text{K}^{-1}$. These give a thermal diffusivity, $\kappa = 1.25 \times 10^{-6}\ \text{m}^2\ \text{s}^{-1}$. The daily temperature variation (period 86,400 s) has $\omega = 7.27 \times 10^{-5}\ \text{s}^{-1}$, so its penetration depth is about 20 cm. The daily variation has negligible effect deeper than about a meter. Similarly, the annual temperature variation accompanying seasonal changes has a penetration depth of 3.8 m and is negligible deeper than about 19 m. Heat-flow measurements made within 20 m of the surface will be contaminated by the daily and annual variations of surface temperature. This is not a problem in the deep oceans, where the Sun’s rays never reach the bottom, but it must be taken into account in continental heat-flow measurements. A serious effect is the role of the ice ages, which recur on a timescale of about 100,000 years and have a penetration depth of several kilometers. The measured temperature gradient must be corrected appropriately.

9.6.2 Cooling of the Oceanic Lithosphere

A thermodynamic problem commonly encountered in geology is the sudden heating or cooling of a body. For example, molten lava intrudes cool host rocks as a dike or sill virtually instantaneously, but the heat is conducted to the adjacent rocks over a long period of time until it slowly dissipates. An important case in the context of plate tectonics is the cooling of the oceanic lithosphere, with its base at the temperature of the hot mantle, T_{m}, and its top in contact with cold ocean-bottom seawater.

Assume that the fresh hot lithosphere is created at a ridge axis as a thin vertical prism with a uniform initial temperature equal to that of the hot mantle, T_{m}. Sea-floor spreading transports the lithosphere horizontally, so, for a constant spreading rate, distance from the ridge axis is proportional to the cooling time (t). Except in the immediate vicinity of a spreading ridge the heat loss may be assumed to be solely in the vertical ($-z$) direction through the surface at $z = 0$, which is in contact with cold ocean-bottom seawater at a temperature of 0°C. Although the lithosphere has finite vertical extent, the temperature in the cooling slab may be approximated as the one-dimensional cooling of a semi-infinite half-space extending to infinity in the z-direction. The error introduced by this simplification is small, and the simple model can be used to estimate the heat flow from the cooling lithosphere with acceptable accuracy.

Box 9.3 The Error Function

Errors may be of two types: systematic and random. A systematic error results, for example, when the measuring device is wrongly calibrated (e.g., if times are measured with a clock that runs “fast” or “slow”). Random errors occur naturally when a value is measured many times. The observations will be distributed randomly about their mean value. Usually there will be a few large deviations from the mean and a lot of small deviations, and there will be as many negative as positive deviations. The scatter of the results can be described by the standard deviation σ of the measurements. Random errors are described by the *normal distribution*, which is often called a "bell curve" because of its shape (Fig. B9.3.1a). If the mean of the distribution of the parameter u is 0 and the standard deviation of the mean is σ, the normal distribution is described by the *probability density function*,

$$f(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(u/\sigma)^2}{2}}.$$

(1)

The *standard normal distribution* is defined so that it has a mean of zero and standard deviation $\sigma = 1$. When integrated from – to +, the area under this curve is

$$\int_{-\infty}^{\infty} f(u)\mathrm{d}u = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} \mathrm{d}u = 1.$$

(2)

The *error function* is closely related to the standard normal distribution. However, only positive values are considered, so the graph of the defining function is similar to the right half of the curve in Fig. B9.3.1a. It is given as

$$f(u) = \frac{2}{\sqrt{\pi}} e^{-u^2}.$$

(3)

The area under this curve from the origin at $u = 0$ to the value $u = x$ (Fig. B9.3.1b) defines the *error function* erf(x):

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2}\,\mathrm{d}u.$$

(4)

The complementary error function, erfc(x), is defined as

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-u^2}\,\mathrm{d}u.$$

(5)

The value of erf(x) or erfc(x) for any particular value of x may be obtained from tables. Some useful properties of the error function and complementary error function are:

$$\operatorname{erf}(\infty) = \frac{2}{\sqrt{\pi}} \int_0^\infty e^{-u^2}\,\mathrm{d}u = 1,$$

(6)

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\operatorname{erf}(x)\right) = \frac{2}{\sqrt{\pi}} e^{-x^2},$$

(7)

$$\begin{aligned}\int_x^{\infty}\operatorname{erfc}(x)\,\mathrm{d}x &= [x\operatorname{erfc}(x)]_x^{\infty} - \int_x^{\infty} x\,\frac{\mathrm{d}}{\mathrm{d}x}\Big(\operatorname{erfc}(x)\Big)\,\mathrm{d}x \\ &= -x\operatorname{erfc}(x) - \int_x^{\infty} x\left(-\frac{2}{\sqrt{\pi}}e^{-x^2}\right)\mathrm{d}x \\ &= \frac{2}{\sqrt{\pi}}\int_x^{\infty} xe^{-x^2}\,\mathrm{d}x - x\operatorname{erfc}(x) \\ &= \frac{1}{\sqrt{\pi}}[-e^{-x^2}]_x^{\infty} - x\operatorname{erfc}(x) \\ &= \frac{e^{-x^2}}{\sqrt{\pi}} - x\operatorname{erfc}(x),\end{aligned}$$

(8)

$$\int_0^{\infty}\operatorname{erfc}(x)\,\mathrm{d}x = \frac{1}{\sqrt{\pi}}.$$

(9)

Some values of erf(*x*) are listed in the following table.

x	**erf(*x*)**	***x***	**erf(*x*)**	***x***	**erf(*x*)**	***x***	**erf(*x*)**
0.05	0.05637	0.3	0.32863	0.6	0.60386	1.2	0.91031
0.1	0.11246	0.35	0.37938	0.7	0.67780	1.4	0.95229
0.15	0.16800	0.4	0.42839	0.8	0.74210	1.6	0.97635
0.2	0.22270	0.45	0.47548	0.9	0.79691	1.8	0.98909
0.25	0.27633	0.5	0.52050	1.0	0.84270	2.0	0.99532

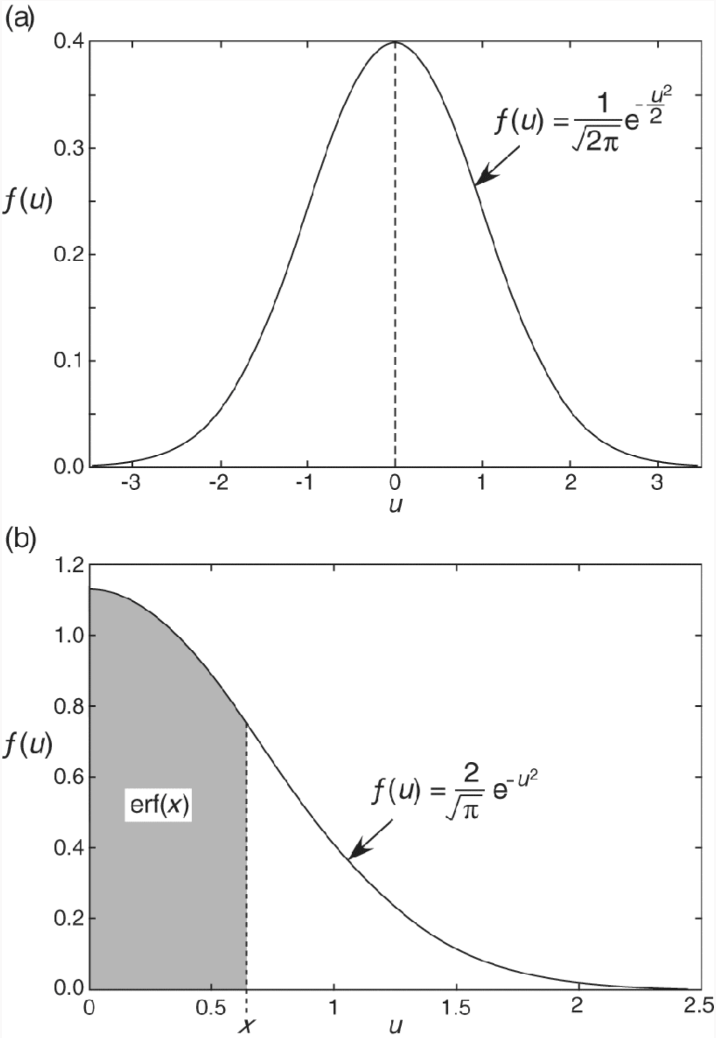


Fig. B9.3.1 (a) The normal distribution, and (b) the error function.

The one-dimensional cooling of a semi-infinite half-space is described in Appendix B. The solution with the stated boundary conditions involves the error function, which is described in Box 9.3. The error function (erf) and complementary error function (erfc) of the parameter *x* are defined as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-u^2}\,du$$

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x).$$

(9.33)

The shapes of these functions are shown in Fig. 9.7; their values for any particular value of x are obtained from tables, just like other statistical or trigonometric functions. The complementary error function sinks asymptotically to zero, and is effectively zero for $x \geq 2$. The temperature T at depth z and time t after the half-space starts to cool is given by

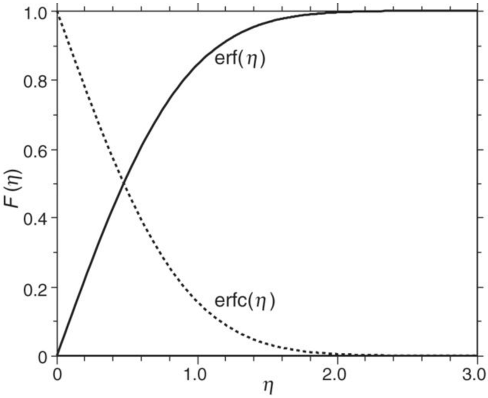


Fig. 9.7 The error function erf(η) and complementary error function erfc(η).

$$T = T_m \operatorname{erf}\left(\frac{z}{2\sqrt{\kappa t}}\right) = T_m \operatorname{erf}(\eta) \quad \text{with} \quad \eta = \frac{z}{2\sqrt{\kappa t}},$$

(9.34)

where T_m is the temperature of the hot mantle, κ is the thermal diffusivity of the half-space, and the surface ($z = 0$) is at the temperature of the ocean floor, which is taken to be 0°C.

The heat loss from the cooling lithosphere can be computed from the temperature distribution in Eq. (9.34). The heat flow is proportional to the temperature gradient (Eq. 9.15). The heat flowing out of a semi-infinite half-space is thus obtained by differentiating Eq. (9.34) with respect to z:

$$q_z = -k\frac{\partial T}{\partial z} = -k\frac{\partial \eta}{\partial z}\frac{\partial T}{\partial \eta}$$
$$= -k\frac{z}{2\sqrt{\kappa t}}\frac{\partial}{\partial \eta}\{T_{\mathrm{m}}\mathrm{erf}(\eta)\}$$
$$= -\frac{kT_{\mathrm{m}}}{2\sqrt{\kappa t}}\frac{\partial}{\partial \eta}\left\{\frac{2}{\sqrt{\pi}}\int_0^{\eta} \mathrm{e}^{-u^2}\mathrm{d}u\right\},$$

(9.35)

which simplifies to

$$q_z = -\frac{kT_{\mathrm{m}}}{\sqrt{\pi \kappa t}}\mathrm{e}^{-\eta^2}.$$

(9.36)

At the surface, $z = 0$, $\eta = 0$, and $\exp(-\eta^2) = 1$. The surface heat flow at time t is

$$q_z = -\frac{kT_{\mathrm{m}}}{\sqrt{\pi \kappa t}}.$$

(9.37)

The negative sign here indicates that the heat flows upward, in the direction of decreasing z. The semi-infinite half-space is quite a good model for the cooling of oceanic lithosphere. The oceanic heat flow indeed varies with distance from an oceanic ridge as $1/\sqrt{t}$, where t is the corresponding age of the lithosphere. Models for the cooling of oceanic lithosphere are discussed further in Section 9.8.3.

9.7 Continental Heat Flow

The computation of heat flow at a locality requires two measurements. The thermal conductivities of a representative suite of samples of the local rocks are measured in the laboratory. The temperature gradient is measured in the field at the investigation site. At continental sites this is usually carried out in a borehole (Fig. 9.8). There are several ways of determining the temperature in the borehole. During commercial drilling the temperature of the drilling fluid can be measured as it returns to the surface. This gives a more or less continuous record, but is influenced strongly by the heat generated during the drilling. At times when drilling is interrupted, the bottom-hole temperature can be measured. Both of these methods give data of possibly commercial interest but they are too inaccurate for heat-flow determination.

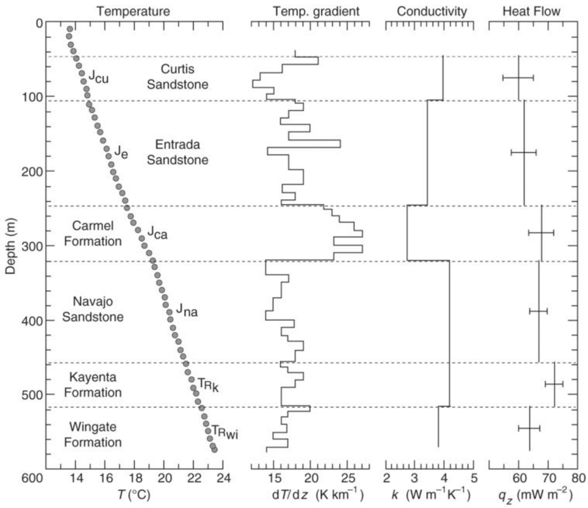


Fig. 9.8 Computation of heat flow by the interval method for geothermal data from drillhole WSR.1 on the Colorado Plateau of the western USA; horizontal bars show standard deviations of measurements in each depth interval.

Source: W.G. Powell, D.S. Chapman, N. Balling and A.E. Beck, Continental heat-flow density, in: *Handbook of Terrestrial Heat-Flow Density Determinations*, R. Haenel, L. Rybach and L. Stegena, eds., pp. 167–222, Dordrecht: Kluwer Academic Publishers, 1988.

In-hole measurements of temperature for heat-flow analyses are made by lowering a temperature-logging tool into the borehole and continuously logging the temperature during its descent. The circulation of drilling fluids redistributes heat in the hole, so it is necessary to allow some time after drilling has ceased for the hole to return to thermal equilibrium with the penetrated formations. The temperature of the water in the hole is taken to be the ambient temperature of the adjacent rocks, provided there are no convection currents.

The most common devices for measuring temperature are the platinum resistance thermometer and the thermistor. A thermistor is a ceramic solid-state device with an electrical resistance that is strongly dependent on temperature. Its resistance depends nonlinearly on temperature, requiring accurate calibration, but the sensitivity of the device makes feasible the measurement of temperature differences of 0.001–0.01 K. The platinum resistance thermometer and the thermistor are used in two basic ways. In one method the sensor element constitutes an arm of a sensitive Wheatstone bridge, with which its resistance is measured directly. The other common method uses the thermal sensor as the resistive element in a tuned electrical circuit. The tuned frequency depends on the resistance of the sensor element, which is related in a known way to temperature.

From the measured temperature distribution, the average temperature gradient is computed for a geological unit or a selected depth interval (Fig. 9.8). The gradient is then multiplied by the mean thermal conductivity of the rocks to obtain the interval (or formation) heat flow. Thermal conductivity can show large variations even between adjacent samples, so the *harmonic mean* of at least four samples is usually used. The interval heat-flow values may then be averaged to obtain the mean heat flow for the borehole (Table 9.4).

Table 9.4 Computation of heat flow from temperature measurements in WSR.1, a 570 m deep borehole, and thermal conductivity measurements on cored samples

Depth interval (m)	**Temperature gradient (K km^{-1})**	**Thermal conductivity (W m^{-1} K^{-1})**	**Interval heat flow (mW m^{-2})**
45–105	15.0	3.96	60
105–245	18.0	3.43	62
245–320	24.8	2.75	68
320–455	16.0	4.18	67
455–515	17.2	4.20	72
515–575	16.5	3.86	64
Mean heat flow			*65*

Source: after Powell et al. (1988).

The computation of continental heat flow from borehole data requires the implementation of several corrections. An important assumption is that heat flow is only vertical. Well below the surface the isotherms (surfaces of constant temperature) are flat-lying and the flow of heat (normal to the isotherms) is vertical. However, the surface of the Earth is also presumed to be locally isothermal (i.e., to have constant temperature) near to the borehole. The near-surface isotherms adapt to the topography (Fig. 9.9) so that the direction of heat flow is deflected and acquires a horizontal component, while the vertical temperature gradient is also modified. Consequently, heat flow measured in a borehole must be corrected for the effects of local topography.

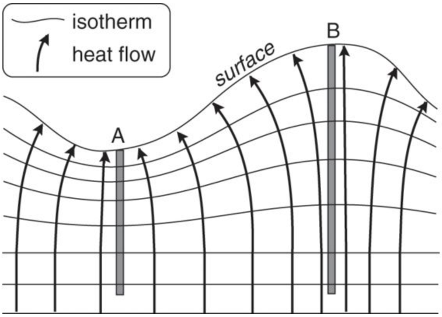


Fig. 9.9 Schematic effect of surface topography on isotherms (solid lines) and the direction of heat flow (dashed lines).

The need for a topographic correction was recognized late in the nineteenth century. Further corrections must be applied for long-term effects such as the penetration of external heat related to cyclical climatic changes, such as the ice ages. Erosion, sedimentation, and changes in the thermal conductivities of surface soils are other long-term effects that may require compensation.

9.7.1 Surface Temperature Changes Inferred from Borehole Temperature Profiles

The main variation of temperature with depth in the Earth is the geothermal gradient related to the outflow of heat from the Earth's deep interior. Over limited depth intervals the temperature profile is linear, its slope varying by region with local conditions. However, changes in temperature at Earth's surface affect the subsurface temperature distribution close to the surface (Section 9.6.1). Rapid temperature changes have shallow penetration, but slow changes can modify subsurface temperatures well below the surface. For example, daily variations do not reach below about a meter, but temperature variations over a century may extend to about 200 m depth.

The temperature profiles in boreholes from eastern Canada (Fig. 9.10) show effects of surface temperature variations above depths of 180–250 m, below which the regular geothermal gradient is evident. The curved upper segment contains information about the history of temperature changes at the surface. This history can be retrieved by inversion of the borehole temperature data, which is a complex and sophisticated process. A large database of borehole temperature measurements, used for the determination of the global heat flow pattern (Section 9.8.2), has been analyzed to obtain the variation of mean surface temperature since 1500 (Fig. 9.11). Instrumental measurements of air temperature since 1860 agree well with the long-term surface temperatures estimated from the borehole data. The results show a consistent increase of surface temperature over the past five centuries.

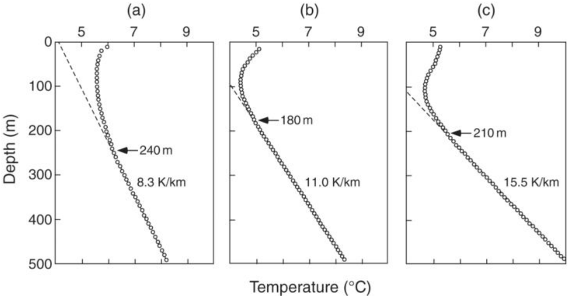


Fig. 9.10 Temperature–depth profiles in three boreholes in eastern Canada. The linear segment in the deeper part of each borehole is the local geothermal gradient. Climatically induced variations in the ground surface temperature result in the curved segment superposed on the linear record in approximately the upper 200 m of each profile.

Source: H.N. Pollack and S. Huang, Climate reconstruction from subsurface temperatures, *Annu Rev. Earth Planet. Sci.*, 28, 339–365, 2000.

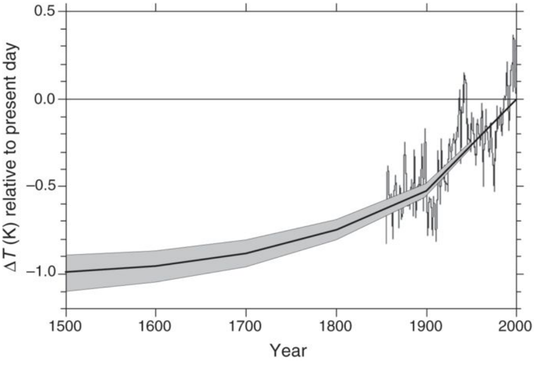


Fig. 9.11 History of surface temperature change (with ±1 standard error, shaded area) inferred for the past 500 years from a global database of borehole temperature measurements. The superposed signal since 1860 is a five-year running mean of the globally averaged instrumental record of surface air temperature.

Source: H.N. Pollack and S. Huang, Climate reconstruction from subsurface temperatures, Annu. Rev. *Earth Planet. Sci.*, 28, 339–65, 2000.

9.7.2 Variation of Continental Heat Flow with Age

Many processes contribute to continental heat flow. Apart from the heat generated by radioactive decay, the most important sources are those related to tectonic events. During an orogenic episode various phenomena may introduce heat into the continental crust. Rocks may be deformed and metamorphosed in areas of continental collision. In extensional regions the crust may be thinned, with intrusion of magma. Uplift and erosion of elevated areas and deposition in sedimentary basins also affect the surface heat flow. After a tectonic event, convective cooling takes place efficiently in circulating fluids, while some excess heat is lost by conductive cooling. Consequently, the variation of continental heat flow with time is best understood in terms of the tectonothermal age, which is the age of the last tectonic or magmatic event at a measurement site. The continental heat-flow values comprise a broad spectrum. Even when grouped in broad age categories there is a large degree of overlap (Fig. 9.12). The greatest scatter is seen in the youngest regions. The mean heat-flow decreases with increasing crustal or tectonothermal age (Fig. 9.13), falling from 70–80 mW m^{-2} in young provinces to a steady-state value of 40–50 mW m^{-2} in Precambrian regions older than ~800 Ma .

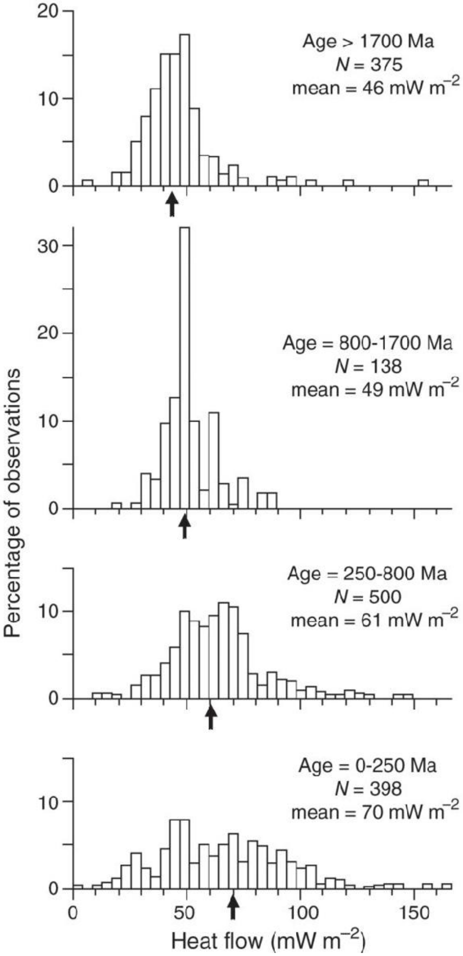


Fig. 9.12 Histograms of continental heat flow for four different age provinces.

Source: J.C. Sclater, B. Parsons and C. Jaupart, Oceans and continents: similarities and differences in the mechanisms of heat loss, *J. Geophys. Res.* 86, 11535–11552, 1981.

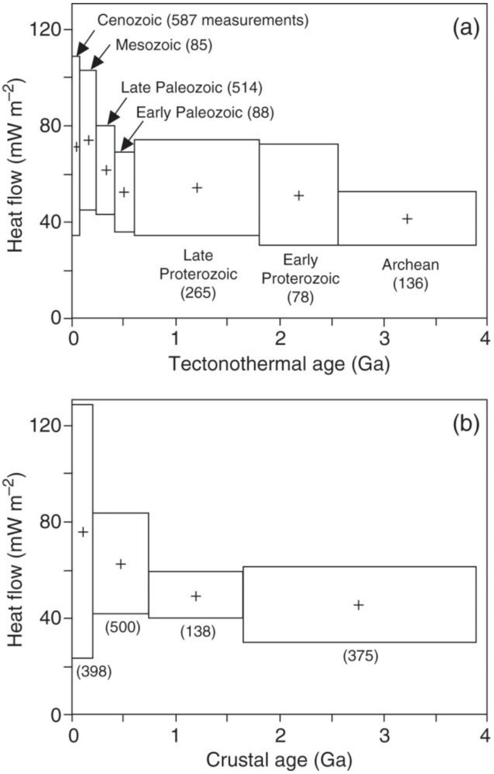


Fig. 9.13 Continental heat-flow data averaged (a) by tectonothermal age, defined as the age of the last major tectonic or magmatic event. Data source: I. Vitorello and H.N. Pollack, On the variation of continental heat flow with age and the thermal evolution of continents, *J. Geophys. Res.* 85, 983–995, 1980. And (b) by radiometric crustal age. Data source: J.C. Sclater, C. Jaupart and D. Galson, The heat flow through oceanic and continental crust and the heat loss of the earth, *Rev. Geophys. Space Phys.* 18, 269–311, 1980. The width of each box shows the age range of the data; the height represents one standard deviation on each side of the mean heat flow indicated by the cross at the center of each box. Numbers indicate the quantity of data for each box.

9.7.3 Heat Transfer through Porous Crustal Rocks

The transfer of heat through the continental crust takes place not only by conduction but also by advection. A sediment or rock is composed of mineral grains closely packed in contact with each other. The pore spaces between the grains can represent an appreciable fraction of the total volume of a rock sample. This fraction, sometimes expressed as a percentage, is the *porosity* ϕ of the rock. A highly porous rock, for example, may have a porosity of 0.3, or 30%, implying that only 70% of the rock is solid mineral. The porosity depends on how the mineral grains are arranged, how well they are cemented, and on their degree of sorting. Well-sorted sediments have fairly uniform grain and pore sizes; in poorly sorted sediments, there is a range of grain sizes, so the finer grains may "block" the voids between the larger grains, reducing the porosity. Igneous and metamorphic rocks often contain cracks and fissures, which, if sufficiently numerous, may give these rocks a low porosity. The degree to which the pore spaces are connected with each other determines the *permeability* of the rock, which is its ability to transmit fluids such as water and petroleum.

Permeability is defined by an empirical relationship observed in 1856 by Henry Darcy, a French hydraulic engineer. He observed that the volume of fluid per second crossing a surface was proportional to the area of the surface (A) and the gradient of the hydraulic pressure head ($\mathrm{d}p/\mathrm{d}x$) driving the flow, assumed here to be in the x-direction, and inversely to the viscosity (μ) of the liquid. If the velocity of the flow is v, then in one second the volume crossing a surface A is (vA), so we have

$$(vA) \propto \frac{A}{\mu}\frac{\mathrm{d}p}{\mathrm{d}x} = -k\frac{A}{\mu}\frac{\mathrm{d}p}{\mathrm{d}x}, \tag{9.38}$$

$$v = -\frac{k}{\mu}\frac{dp}{dx} = -K\frac{dp}{dx}. \tag{9.39}$$

This equation, in which $K = k/\mu$ defines the *hydraulic conductivity*, is known as Darcy's law. Its derivation and form are analogous to the law of heat conduction (Eq. 9.14 and Fig 9.3) and Ohm's law for electrical currents (Eq. 10.7 and Fig. 10.2). The negative sign in the equation indicates that the flow is in the direction of decreasing pressure. The constant k is the permeability, which has dimensions m^2. However, the unit of permeability used in hydraulic engineering is called the *darcy* (equivalent to 9.87×10^{-13} m^2), or, more practically, the *millidarcy* (md). For example, an approximate range for the permeability of gravel is 10^5–10^8 md, that of sandstone is 1–100 md, and that of granite is 10^{-3}–10^{-5} md.

The ability of fluids to flow through crustal rocks enables them to transmit heat. In this case the process of heat transfer is not by convection, because the fluid motion is not driven by temperature differences but by the pressure gradient. The heat transfer "piggybacks" on the fluid motion, and the process is called *advection* (Section 9.4.2). The motion of water through the continental crust provides a source of geothermal energy, which can be tapped in several interesting ways for commercial purposes.

9.8 Oceanic Heat Flow

Whereas the mean altitude of the continents is only 840 m above sea level, the mean depth of the oceans is 3880 m; the abyssal plains in large ocean basins are 5–6 km deep. At these depths extra-terrestrial heat sources have no effect on heat-flow measurements, and the flatness of the ocean bottom (except near ridge systems or seamounts) obviates the need for topographic corrections. Measuring the heat flow through the ocean bottom presents technical difficulties that were overcome with the development of the Ewing piston corer. This device, intended for taking long cores of marine sediment from the ocean floor, enables *in-situ* measurement of the temperature gradient. It consists of a heavily weighted, hollow sampling pipe (Fig. 9.14a), commonly about 10 m long, although in special cases cores over 20 m in length have been taken (very long coring pipes tend to bend before they reach maximum penetration). A plunger inside the pipe is displaced by sediment during coring and makes a seal with the sediment surface, so that sample loss and core deformation are minimized when the core is withdrawn from the ocean floor. Thermistors are mounted on short arms a few centimeters from the body of the pipe, and the temperatures are recorded in a water-tight casement. The instrument is lowered from a surface ship until a free-dangling trigger-weight makes contact with the bottom (Fig. 9.14b). This releases the corer, which falls freely and is driven into the sediment by the one-ton lead weight. The friction accompanying this process generates heat, but the ambient temperatures in the sediments can be measured and recorded before the heat reaches the offset sensors (Fig. 9.14c). The sediment-filled corer is hauled back on board the ship. The recovered core is used for paleontological, sedimentological, geochemical, magnetostratigraphic, and other scientific analyses.

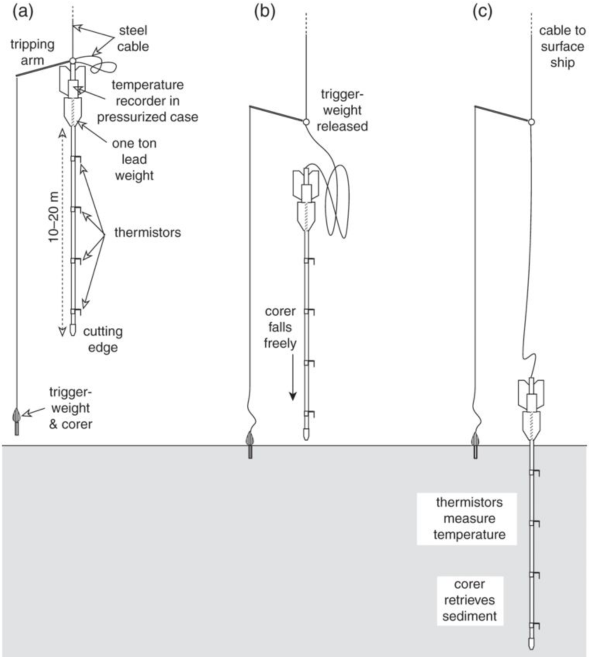

(a)
tripping arm
steel cable
temperature recorder in pressurized case
one ton lead weight
10–20 m
thermistors
cutting edge
trigger-weight & corer
(b)
trigger-weight released
corer falls freely
(c)
cable to surface ship
thermistors measure temperature
corer retrieves sediment

Fig. 9.14 Method of measuring oceanic heat flow and recovering samples of marine sediments: (a) a coring device is lowered by cable to the sea floor; (b) when a trigger-weight contacts the bottom, the corer falls freely; and (c) temperature measurements are made in the ocean floor and the sediment-filled corer is recovered to the surface ship.

Special probes have been devised explicitly for *in-situ* measurement of heat flow. They consist of two parallel tubes about 3–10 m in length and 5 cm apart. One tube is about 5–10 cm in diameter and provides strength; the other, about 1 cm in diameter, is oil-filled and contains arrays of thermistors. After penetration of the ocean-bottom sediments, as described for the Ewing piston corer, the equilibrium temperature gradient is measured. A known electrical current, either constant in value or pulsed, is then passed along a heating wire and the temperature response is recorded. The observations allow the thermal conductivity of the sediment to be found. In this way a complete determination of heat flow is obtained without having to recover the contents of the corer.

9.8.1 Variation of Oceanic Heat Flow and Depth with Lithospheric Age

The most striking feature of oceanic heat flow is the strong relationship between the heat flow and distance from the axis of an oceanic ridge. The heat flow is highest near the ridge axis and decreases with increasing distance from it. For a uniform sea-floor spreading rate the age of the oceanic crust (and lithosphere) is proportional to the distance from the ridge axis, and so the heat flow decreases with increasing age (Fig. 9.15). The lithospheric plate accretes at the spreading center, and as the hot material is transported away from the ridge crest it gradually cools. Model calculations for the temperature in the cooling plate are discussed in the next section; they all predict that the heat flow q caused by cooling of the plate decreases with age t as $1/\sqrt{t}$, when the age of the plate is less than about 55–70 Ma. Older lithosphere cools slightly less rapidly. Currently the decrease of heat flow with age is best explained by a global model called the global depth and heat flow model (GDH1). The model predicts the following relationships between heat flow (q, mW m^{-2}) and age (t, Ma):

$$q = \frac{510}{\sqrt{t}} \qquad (t \leq 55\ \text{Ma})$$

$$\begin{aligned} q &= q_s\,[1 + 2\exp(-\kappa\pi^2 t/a^2)] \qquad (t > 55\ \text{Ma}) \\ &= 48 + 96\exp(-0.0278t). \end{aligned} \qquad (9.40)$$

Here, q_s is the asymptotic heat flow, to which the heat flow decreases over very old oceanic crust ($\approx$ 48 mW m^{-2}), a is the asymptotic thickness of old oceanic lithosphere ($\approx$ 95 km), and κ is its thermal diffusivity ($\approx 0.8 \times 10^{-6}$ m^2 s^{-1}).

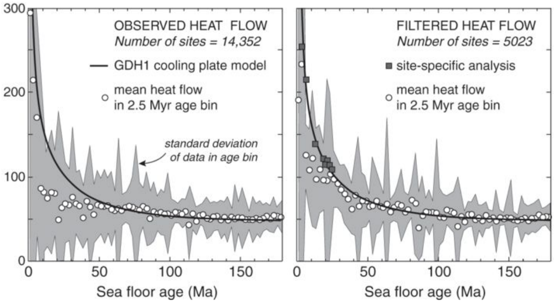


Fig. 9.15 Variations of oceanic heat flow with lithospheric age. Observed data are averaged in 2.5 Myr bins, and compared to the heat flow predicted by the plate cooling model GDH1 (Stein and Stein, 1992). (a) Global data from all the oceans, showing low values near the ocean ridges due to the influence of hydrothermal circulation. (b) Data from selected sites remote from seamounts, corrected for sedimentation effects; squares mark sites on young sea floor where more detailed analysis of hydrothermal circulation was made.

Source: Hasterok, et al., Oceanic heat flow: implications for global heat loss, *Earth Planet. Sci. Lett.*, 311, 386–395, 2011

Close to a ridge axis the measured heat flow is unpredictable: extremely high values and very low values have been recorded. Over young lithosphere the observed heat flow is systematically less than the values predicted by cooling models (Fig. 9.15a). The discrepancy is a result of the process of accretion of the new lithosphere. At a ridge crest, magma erupts in a narrow zone through feeder dikes and/or supplies horizontal lava flows. Very hot material is brought into contact with seawater, which cools and fractures the fresh rock. The water is in turn heated rapidly and a hydrothermal circulation is set up in the cracks and fissures, which transports heat out of the lithosphere by convection. The eruption of hot hydrothermal currents has been observed directly from manned submersibles in the axial zones of oceanic ridges. Scientists on these deep-diving expeditions witnessed strong outpourings of mineral-rich hot water (called “black smokers” and “white smokers”) in the narrow axial rift valley. The heat output of these vents is high: the power associated with a single vent has been estimated to be about 200 MW.

About 30% of the hydrothermal circulation takes place very near to the ridge axis through crust younger than 1 Ma. The rest is due to off-ridge circulation, which is possible because the fractured crust is still permeable to seawater at large distances from the ridge axis. As the young crust moves away from the ridge, sedimentation covers the igneous basement with a progressively thicker layer of low-permeability sediments, inhibiting the convective heat loss. The hydrothermal circulation eventually ceases, partly because it is sealed by the thick sediment cover, but also because the cracks and pore spaces in the crust become closed with increasing age. This is estimated to have taken place in about 55–70 Ma, because for greater ages the observed decrease of heat flow with age is close to that predicted by plate cooling models. The hydrothermal circulation in oceanic crust is an important part of the Earth’s heat loss. It accounts for about one-third of the total oceanic heat flow, and one-quarter of the global heat flow.

The free-air gravity anomaly over an oceanic ridge system is generally small and related to ocean-bottom topography (Section 4.2.4.3), which suggests that the ridge system is isostatically compensated. As hot material injected at the ridge crest cools, its volume contracts, and its density increases. To maintain isostatic equilibrium, a vertical column sinks into the supporting substratum as it cools. Consequently, the depth of the ocean floor (the top surface of the column) is expected to increase with age of the lithosphere. The cooling half-space model predicts an increase in depth proportional to $\sqrt{t}$, where t is the age of the lithosphere, and this is observed up to an age of about 80 Ma (Fig. 9.16). However, the square root relationship is not the best fit to the observations. Other cooling models fit the observations more satisfactorily, although the differences from one model to another are small. Beyond 20 Ma the data are better fitted by an exponential decay. The optimum relationships between depth (d, m) and age (t, Ma) can be written as

$$\begin{aligned} d &= 2600 + 365\sqrt{t} && (t < 20\ \mathrm{Ma}) \\ d &= d_r + d_s[1 - (8/\pi^2)\exp(-\kappa\pi^2 t/a^2)] && (t \geq 20\ \mathrm{Ma}) \\ &= 5651 - 2473\exp(-0.0278t), \end{aligned} \tag{9.41}$$

where d_r is the mean depth of the ocean floor at ridge crests, d_s is the asymptotic subsidence of old lithosphere, and the other parameters are as before.

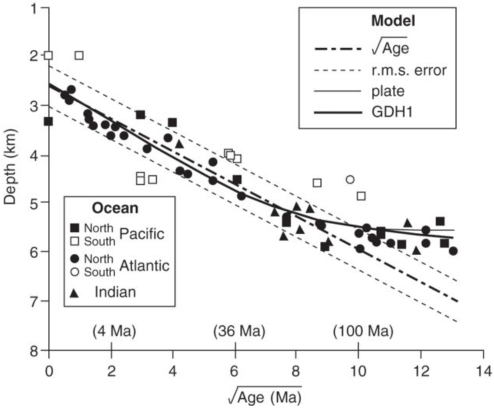


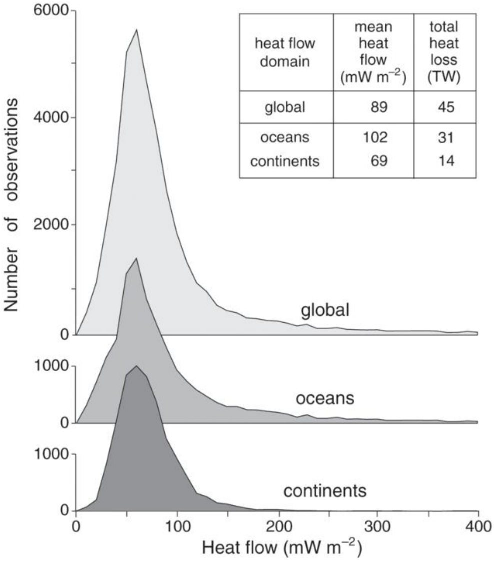
Fig. 9.16 Relationship between mean ocean depth and the square root of age for the Atlantic, Pacific, and Indian Oceans, compared with theoretical curves for different models of plate structure.

Source: H.P. Johnson and R.L. Carlson, Variation of sea floor depth with age: A test of models based on drilling results, *Geophys. Res. Lett.* 19, 1971–1974, 1992.

9.8.2 Global Heat Flow

Oceanic heat flow has been measured routinely in oceanographic surveys since the 1950s and *in-situ* profiles have been made since the 1970s. In contrast to the measurement of continental heat flow, it is not necessary to have an available (and usually expensive) drillhole. However, the areal coverage of the oceans by heat-flow measurements is uneven. Large areas in the Pacific Ocean are still unsurveyed, and most of the oceanic areas south of about latitude 35°S (the approximate latitude of Cape Town or Buenos Aires) are unsurveyed or only sparsely covered. The uneven data distribution is dense along the tracks of research vessels and absent or meager between them. The sites of measured continental heat flow are even more irregularly distributed. Antarctica, most of the interiors of Africa and South America, and large expanses of Asia are either devoid of heat-flow data or are represented by only a few sites.

In order to compensate for the uneven coverage, digital maps of the Earth's geology were assigned representative heat flows according to age and rock type. The geographical information system (GIS) defines the geological units as irregular polygons on the globe's surface, so that the heat flow for a specific geology can be mapped. This resulted in a global dataset of more than 38,000 heat-flow values, comprising more than 16,000 on land and more than 22,000 in the oceans. The heat-flow values are distributed over a wide range for each domain (Fig. 9.17) and have similar characteristics, extending from very low values to more than 300 mW m^{-2}. On the continents (Fig. 9.13) and also in the oceans (Fig. 9.15), the heat flow varies with crustal age. High values on the continents are found in volcanic and tectonically active regions, while the highest values in the oceans occur at the axes of oceanic ridges. Low values near to oceanic ridges are explained by heat loss due to hydrothermal circulation instead of conduction.



heat flow domain	mean heat flow (mW m^{-2})	total heat loss (TW)
global	89	45
oceans	102	31
continents	69	14

Fig. 9.17 Distributions of continental, oceanic, and global heat-flow values from 38,347 observations. Oceanic heat flow younger than 65 Ma has been corrected for the loss of heat by hydrothermal circulation.

Source: J.H. Davies & D.R. Davies, Earth's surface heat flux, *Solid Earth*, 1, 5–24, 2010.

The oceanic data must be corrected for this loss, generally assuming a model of ocean plate cooling (as described in the next section). To determine global heat-flow statistics, the area of each measurement polygon is multiplied by the mean heat flow measured or estimated for the polygon. This gives the rate of loss of heat from a polygon. Combining polygons gives the heat flow by age for oceanic and continental domains (Table 9.5). This procedure makes it possible to estimate heat flow for regions that are poorly represented by measurement sites. The computations give a mean heat flow of 69.4 mW m^{-2} for the continental regions, and 102.3 mW m^{-2} for the oceanic dataset; values on ocean ridges average 133 mW m^{-2}, whereas old oceanic lithosphere has cooled to around 70 mW m^{-2}. The global mean heat flow, weighted by area, is 89 mW m^{-2}. The Earth's surface area is 5.101 × 10^8 km^2, thus the estimated rate of global heat loss is found to be 45.4 TW (4.54×10^{13} W, equivalent to an annual heat loss of 1.4×10^{21} J). Roughly 70% of the heat is lost through the oceans and 30% through the continents.

Table 9.5 Mean heat-flow values for the oceans and continents, based on measurements at 38,040 sites
Heat flow and heat loss in italics are corrected for hydrothermal circulation in young oceanic crust (Stein and Stein, 1992). One terawatt (TW) = 10^6 MW = 10^{12} W.

Description	Number of sites	Area (10^6 km^2)	Heat flow (mW m^{-2})	Heat loss (TW)
Oceanic areas				
Quaternary and Pliocene	4397	16.1	*424.5*	*6.85*
Neogene (undifferentiated)	29	0.5	*204.5*	*0.09*
Miocene	3408	50.3	*137.9*	*6.94*
Oligocene	1063	31.7	*92.3*	*2.93*
Eocene	988	37.6	*73.7*	*2.77*
Paleocene	652	21.1	*63.0*	*1.33*
Upper Cretaceous	1571	45.5	66.8	3.04
Lower Cretaceous	1299	34.6	61.1	2.11
Cretaceous (undifferentiated)	58	1.2	54.7	0.06
Upper–Middle Jurassic	1107	17.7	54.0	0.96
Cretaceous/ Jurassic (undifferentiated)	224	0.6	34.2	0.02
Age unknown	642	24.1	71.5	2.27
Seamounts	993	21.8	73.8	1.61
Oceanic summary	**16,531**	**303**	**102.3**	**31.0**
Continental areas				
Cenozoic	9726	68.7	75.9	5.21

Mesozoic	2967	26.3	66.7	1.75
Paleozoic	2225	22.9	58.9	1.35
Proterozoic	1563	28.7	59.9	1.72
Continental and island arc margins	4556	58.7	73.8	4.24
Other	472	1.7	54.5	0.09
Continental summary	**21,509**	**207**	**69.4**	**14.3**
Global summary	**38,040**	**510**	**89.0**	**45.3**

(*data source:* Davies and Davies, 2010)

The relationships between heat-flow values, crustal age, and geological characteristics makes it possible to create a map of global heat flow that allows for the uneven distribution of actual measurements. The gridded data are then fitted by spherical harmonic functions (as in the representation of the geoid, Section 3.4.5), producing smooth contours of equal heat flow (Fig. 9.18). If the global mean value is subtracted, the Earth's surface can be divided into regions with above-average and below-average heat flow, respectively (Fig. 9.19). The regions with above-average heat flow are notably associated with the oceanic ridge systems. About half of the Earth's heat is lost by the cooling of oceanic lithosphere of Cenozoic age (i.e., younger than 66 Ma).

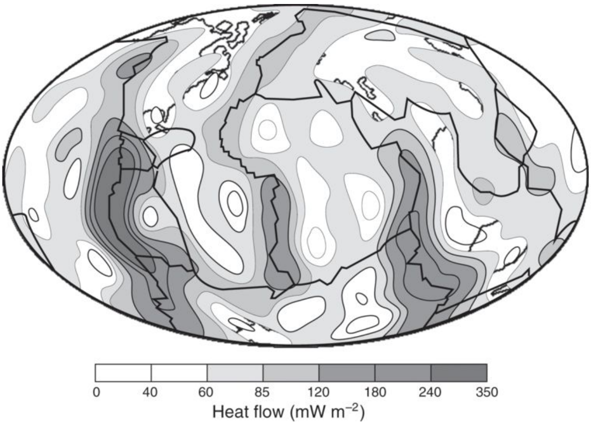


Fig. 9.18 Global distribution of heat flow (mW m–2). The contours show a degree and order 12 spherical harmonic representation of the global heat flow based on direct measurements and empirical estimators for regions without data.

Source: H.N. Pollack, S.J. Hurter and J.R. Johnson, Heat flow from the Earth's interior: analysis of the global data set, *Rev. Geophys.* 31, 267–280, 1993.

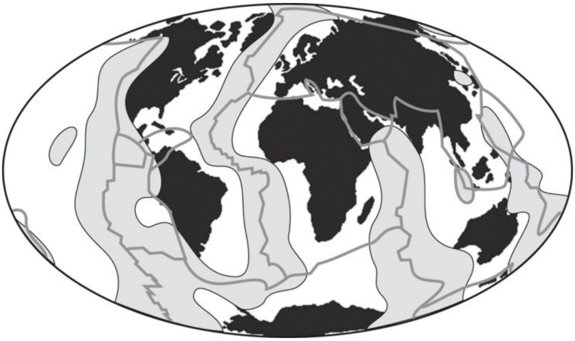


Fig. 9.19 Geographic regions where the heat flow is higher (*lighter-shaded*) and lower (*unshaded*) than the global mean heat flow; lines (*darker shaded*) mark positions of plate boundaries.

Source: H.N. Pollack, S.J. Hurter and J.R. Johnson, Heat flow from the Earth's interior: analysis of the global data set, *Rev. Geophys.* 31, 267–280, 1993

One must keep in mind that a global heat-flow model is based on a mixture of actual heat-flow measurements in regions where they are available, and estimated values for inaccessible regions. Moreover, the measured data near ocean ridges are replaced with values predicted by cooling models to compensate for the known loss of heat by hydrothermal circulation. Nevertheless, global heat-flow maps are the best available representation of the geographical pattern and flux of the heat flowing out of the Earth's interior. Although details may eventually need modification, the main features are not in doubt.

9.8.3 Models for the Cooling of Oceanic Lithosphere

The variations of heat flow and ocean depth with time constrain the possible thermal models for cooling of the lithosphere in different ways. The predicted heat flow is contingent on the temperature gradient in a model, but the oceanic depth is defined by the vertical distribution of density, which, in turn, depends on the volume coefficient of expansion and the temperature profile in the plate. Thus, oceanic bathymetry depends on the temperature integrated over depth.

Several cooling models have been proposed, all of which satisfy the decrease in heat flow and increase in ocean depth with age. The simplest model represents the cooling lithosphere as a semi-infinite half-space (Fig. 9.20a). Initially, the temperature inside the half-space is uniform and higher than on its upper surface, which is maintained at the temperature of cold ocean-bottom seawater. As long as the lithosphere is thin – which it is near the ridge – horizontal heat conduction can be neglected. The heat flow in the uniform half-space is vertical, along the z-axis, and is equivalent to the one-dimensional flow in a thin vertical column (Fig. 9.21a). The spreading process can be envisioned as transporting the column away from the ridge axis, during which conductive cooling takes place and the temperature distribution in the column changes.

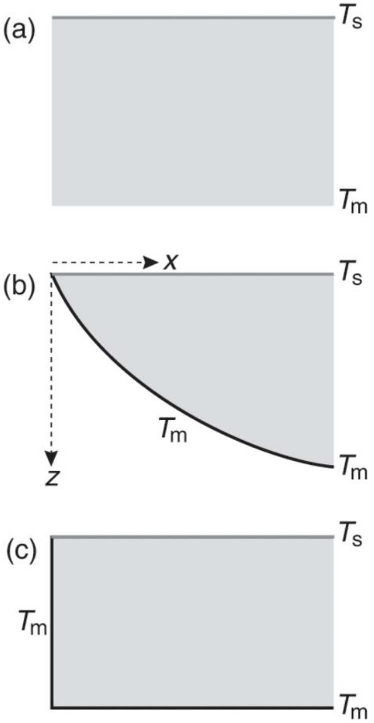

(a)
Ts
Tm
x
(b)
Ts
Tm
z
Tm
(c)
Ts
Tm
Tm

Fig. 9.20 (a) Semi-infinite half-space; (b) thermal boundary layer; and (c) plate models for the cooling of oceanic lithosphere. T_s and T_m are surface and mantle temperature, respectively.

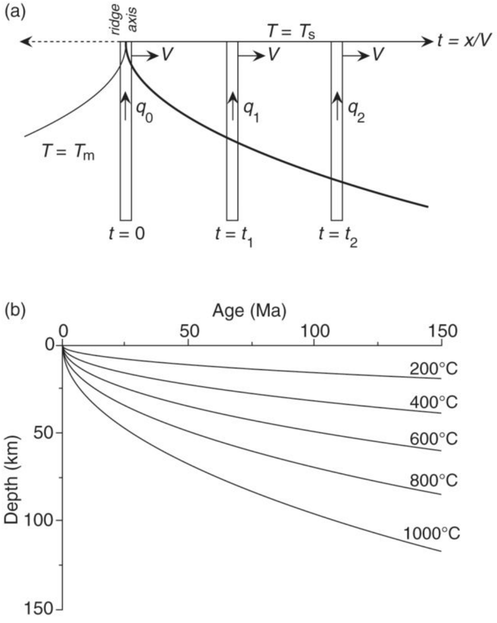


Fig. 9.21 Application of the infinite half-space model to explain the cooling of oceanic lithosphere: (a) vertical heat flow in narrow columns that move away from the ridge crest; and (b) predicted thermal structure in the cooling plate.

Source: D.L. Turcotte and G. Schubert, *Geodynamics: Applications of Continuum Physics to Geological Problems*, 450 pp., New York: John Wiley, 1982.

This model allows us to compute the temperature distribution in the oceanic lithosphere. We need to compute the depth z at which a given temperature T is reached after time t, when the vertical column has moved at velocity V to a distance Vt from the ridge. First, the desired temperature T is expressed as a fraction of the mantle temperature T_m. Using Eq. (9.34) and the appropriate table, the argument η_0 is found, which gives an error function equal to (T/T_m). Setting the numerical value η_0 equal to $z/2\sqrt{t}$ gives the shape of the isotherm for the temperature T:

$$\eta_0 = \frac{z}{2\sqrt{\kappa t}}$$

$$z = (2\eta_0\sqrt{\kappa})\sqrt{t}. \tag{9.42}$$

The isotherms in the cooling lithosphere have a parabolic shape with respect to the time (or horizontal distance) axis, of which only the part for $z > 0$ is of interest. The surface heat flow for this model is given by Eq. (9.37), and so is inversely proportional to $\sqrt{t}$.

The half-space model has some unrealistic aspects. It predicts infinitely large heat flow at the ridge axis, and the initial mantle temperature T_m is approached asymptotically and is only reached at infinite depth. The distances between successive isotherms for equal increments in temperature get progressively larger. The near-surface layer in which the temperature changes are significant has been called a *thermal boundary layer*. Its base is defined arbitrarily as the depth at which the temperature reaches a chosen fraction of T_m. The layer can be regarded as a thermal model of the lithosphere (Fig. 9.20b). Instead of being defined mechanically as the depth where seismic shear waves are attenuated, the base of the lithosphere in the thermal model is an isotherm (Fig. 9.21b). The model predicts that the lithosphere becomes thicker with increasing age, as also inferred from seismic data, and the thickness is proportional to $\sqrt{t}$. As it cools and thickens, the lithosphere sinks deeper into the asthenosphere, so that the ocean depth increases away from a ridge axis. Together with Pratt-type, thermally influenced isostasy the half-space model of lithosphere cooling (Box 9.4) predicts a depth increase that is also proportional to $\sqrt{t}$.

Box 9.4 Variation of Ocean Depth with Age

As oceanic lithosphere moves away from the ridge it cools, thickens, and becomes denser (Fig. 9.21). It sinks progressively into the underlying asthenosphere with time, so that the depth of an oceanic basin increases with age t of the oceanic lithosphere. A simple model accounts for the depth change w by assuming that the lithosphere and asthenosphere are in Pratt-type isostatic balance. The isostatic model is sometimes referred to as *thermal isostasy*.

Compare the composition of two vertical columns of unit cross-sectional area above a compensation level in the asthenosphere at depth D (Fig. B9.4.1). The column below R on the ridge axis consists of hot asthenosphere, of assumed constant density ρ_a and temperature T_a, and the depth d_r of seawater (density ρ_w) above the ridge. The column below B over the adjacent ocean basin is a section through oceanic lithosphere. The density ρ_L of the lithosphere depends on its temperature T_L and the lithosphere thickness L increases with age t. The seawater layer of depth d_r above the ridge is present in both columns, as is the thickness *A* of asthenosphere between the base of the lithosphere and the compensation depth. The isostatic balance is determined by equating the weights of *w* km of seawater and *L* km of lithosphere with the corresponding weight of (*w* + *L*) km of asthenosphere:

$$(w+L)\rho_a g = w\rho_w g + g\int_0^L \rho_L \mathrm{d}z,$$

(1)

$$w(\rho_a - \rho_w) = \int_0^L (\rho_L - \rho_a)\mathrm{d}z.$$

(2)

Thermal isostasy assumes that the lithosphere changes density as it cools. The volume coefficient of expansion *α* is defined by Eq. (9.2) as

$$\alpha = \frac{1}{V}\frac{\partial V}{\partial T} = -\frac{1}{\rho}\frac{\partial \rho}{\partial T},$$

(3)

where density $\rho = M/V$ and thus $dV/V = -d\rho/\rho$. Rewriting Eq. (3), we get

$$\alpha\rho_a(T_a - T_L) = (\rho_L - \rho_a).$$

(4)

The expression for the density difference between the lithosphere and asthenosphere is now substituted into Eq. (2):

$$w = \frac{\alpha \rho_a}{(\rho_a - \rho_w)} \int_0^L (T_a - T_L)\, dz.$$

(5)

The temperature of the lithosphere T_L is given by Eq. (9.34), with T_a instead of T_m. Substituting in Eq. (5), we find

$$\begin{aligned} w &= \frac{\alpha \rho_a}{(\rho_a - \rho_w)} \int_0^L \left(T_a - T_a \operatorname{erf}(\eta)\right) dz \\ &= \frac{\alpha \rho_a T_a}{(\rho_a - \rho_w)} \int_0^L \operatorname{erfc}(\eta) dz, \text{ with } \eta = \frac{z}{2\sqrt{\kappa t}}. \end{aligned}$$

(6)

The complementary error function, erfc(η), decreases almost to zero by η = 2, so the upper limit of integration can be changed from L to ∞ without causing significant error:

$$\begin{aligned} w &= \frac{\alpha \rho_a T_a}{(\rho_a - \rho_w)} \int_0^\infty \operatorname{erfc}(\eta) dz \\ &= \frac{\alpha \rho_a T_a}{(\rho_a - \rho_w)} 2\sqrt{\kappa t} \int_0^\infty \operatorname{erfc}(\eta) d\eta. \end{aligned}$$

(7)

From Eq. 9 in Box 9.3 we have $\int_0^\infty \operatorname{erfc}(\eta) d\eta = \frac{1}{\sqrt{\pi}}$. Thus,

$$w = \frac{2}{\sqrt{\pi}} \frac{\alpha \rho_a T_a}{(\rho_a - \rho_w)} \sqrt{\kappa t}.$$

(8)

This is the amount by which the ocean deepens away from the ridge axis. The total depth of the ocean, taking into account the depth d_r at the ridge, is $d = (d_r + w)$. Optimum values for the parameters of the lithosphere in Eq. (8) are given in the GDH1 model of Stein and Stein (1992): $\alpha = 3.1 \times 10^{-5}\ \mathrm{K}^{-1}$, $\rho_a = 3300\ \mathrm{kg\ m^{-3}}$, $\kappa = (k/c_p\rho_a) = 8.04 \times 10^{-7}\ \mathrm{m^2\ s^{-1}}$. Assuming a mean depth of 2600 m over the ridge axis, a temperature $T_a = 1450°\mathrm{C}$ for the asthenosphere, and $\rho_w = 1030\ \mathrm{kg\ m^{-3}}$ for the density of seawater, the depth of the ocean over crust of age t (in Ma) is given by

$$d = (d_r + w) = 2600 + 370\sqrt{t}. \tag{9}$$

This computed result is close to the depth–age relationship in Eq. (9.39) predicted by the GDH1 model for young lithosphere (age < 20 Ma).

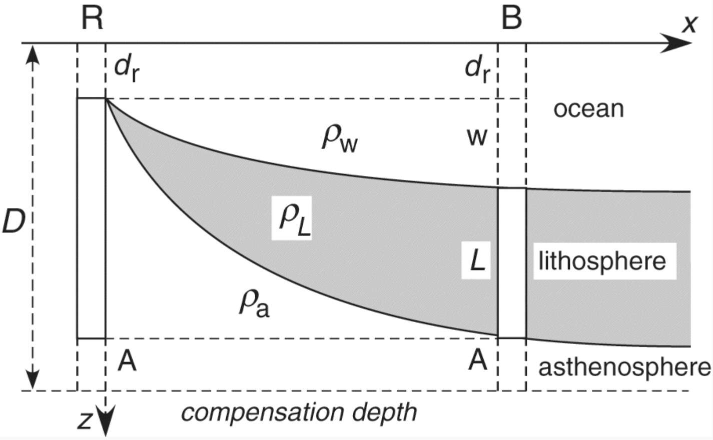


Fig. B9.4.1 Vertical section through oceanic lithosphere from a ridge to an adjacent ocean basin.

Parker and Oldenburg (1973) proposed a modification of the boundary layer model in which a solid lithosphere overlies a fluid asthenosphere. The base of the lithosphere is taken to be the solid–

liquid phase boundary of the material. It is defined by the melting-point isotherm, and denotes a phase change. This is probably a closer representation of the real situation, although by treating the asthenosphere as a fluid, it exaggerates the change in rheology. The temperature of the asthenosphere lies close to the solidus temperature, but its condition is only partially molten (perhaps about 5%).

The half-space and boundary layer models fit the observed variations of heat flow and ocean depth with age for young lithosphere. For ages greater than about 70 Ma the ocean depths in particular are less than predicted by the $\sqrt{t}$ relationship (Fig. 9.16). This suggests that the source of heat from below the lithosphere may be shallower than in the half-space model at large ages. As an alternative to the half-space models the oceanic lithosphere has been modeled as a flat layer or plate of finite thickness, bounded above by cold seawater and with a constant temperature on its lower surface. Far from a ridge axis this model brings hot mantle temperatures nearer to the surface than in the half-space model. Below the ridge axis the vertical edge of the new plate has the same high temperature of its lower surface (Fig. 9.20c), which results in heat being conducted horizontally through the plate. This is not a serious problem as long as the plate is much thinner than its horizontal extent away from a ridge. This condition is clearly met for the main lithospheric plates, which are several thousand kilometers across and only of the order of 100 km thick.

The plate model is not intended to model the vertical mechanical structure of the plate, but only to explain in a phenomenological way the typical age-dependence of both ocean depth and heat flow. The plate thickness in the model is the asymptotic thermal thickness of old oceanic lithosphere and reflects the combined effects of temperature and rheology. Its horizontal isothermal base requires additional deep heat sources that prevent the lithosphere from cooling as a half-space at great ages. The model allows simple computation of the thermal cooling history. The best-known version, proposed by Parsons and

Sclater (1977), assumed a plate thickness of 125 km and a basal temperature of 1350°C. At large distances from each spreading center it gives very good fits to both the observed heat flow (Fig. 9.15) and the ocean depth (Fig. 9.16). The most recent update, the GDH1 plate model, has a thickness of 95 km and a basal temperature of 1450°C. It fits the observations even better.

9.8.4 Thermal Structure of Oceanic Lithosphere

The plate models explain observed thermal data better than the boundary layer models. The boundary layer model is most appropriate near to a ridge axis, and agrees better with other geophysical data, which show that the lithosphere thickens with distance from a ridge. However, the plate model is needed at great distances to explain heat flux and ocean depths over old lithosphere. To reconcile these contrasting attributes, a two-layered model of the lithosphere was proposed (Fig. 9.22). The upper layer is rigid and has a mechanically defined lower boundary, above which heat transfer is by conduction. Below this level the increasing temperature causes a change in mechanical properties. The lower lithosphere is plastic enough to permit material movement, and so behaves like a viscous solid. The base of the upper layer is an isotherm, representing the temperature at which rigidity is lost. The base of the lower lithosphere is a thermally defined boundary, and is also an isotherm. Several suggestions have been made as to how this structure may approximate the plate model for old lithosphere. They include additional heat sources such as radiogenic heating, frictional heating as a result of shear at the base of the lithosphere, and reheating of old lithosphere due to the intrusion of mantle plumes at hotspots. It has also been postulated that, at lithosphere ages greater than about 70 Ma, small-scale convection currents in the lower lithosphere may augment thermal conduction. This would bolster the transfer of heat from the convecting asthenosphere into the lithosphere, effectively giving a thinner lithosphere than in the half-space models. Analysis of the dispersion of seismic surface waves indicates that there are differences in structure between continents and oceans down to about 200 km. This is compatible with the thermal model of a rigid mechanical layer underlain by a convecting thermal boundary layer extending to about 150–200 km.

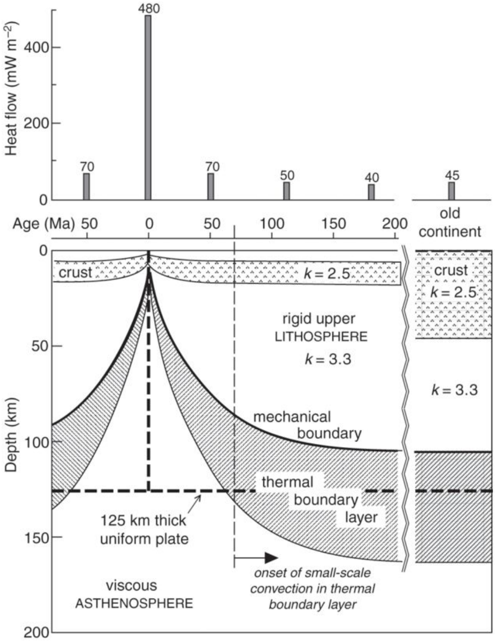

Heat flow (mW m⁻²)
400
200
0
480
70
70
50
40
45
Age (Ma) 50
0
50
100
150
200
old continent
Depth (km)
0
50
100
150
200
crust
k = 2.5
rigid upper LITHOSPHERE
k = 3.3
mechanical boundary
thermal boundary layer
125 km thick uniform plate
viscous ASTHENOSPHERE
onset of small-scale convection in thermal boundary layer
crust
k = 2.5
k = 3.3

Fig. 9.22 Schematic diagram of lithospheric plate structure beneath oceans and continents. The dashed line indicates the approximation as a plate of constant thickness (based upon Parsons and McKenzie, 1978, and Sclater et al., 1981).

9.8.5 Heat Flow at Subduction Zones

The oceanic lithosphere is bent sharply downward beneath the overriding plate in a subduction zone. It extends as an inclined slab deep into the upper mantle, which it penetrates at a rate of a few centimeters per year. The old lithosphere is cold, having lost much of its original heat of formation at the ridge axis. By the time it reaches an ocean trench, the isotherms in the plate are far apart and the temperature gradient is small. The separation of the isotherms is increased by the downward bending of the plate. Since the heat flow is proportional to the temperature gradient, very low heat-flow values ($\approx$ 35 mW m^{-2}) are measured in oceanic trenches. After bending downward, the plate is subducted to great depths, subjecting it to increases in pressure and temperature. Heat is conducted into the plate from the adjacent mantle. This process is so slow that the interior of the subducting slab remains colder than its environment (Fig. 9.23). A temperature of 800°C is normally reached at about 70 km depth in the oceanic plate, but in the descending slab this temperature exists to deeper than 500 km. Above this depth the coldest part of the slab has a horizontal temperature deficit of 800–1000 K.

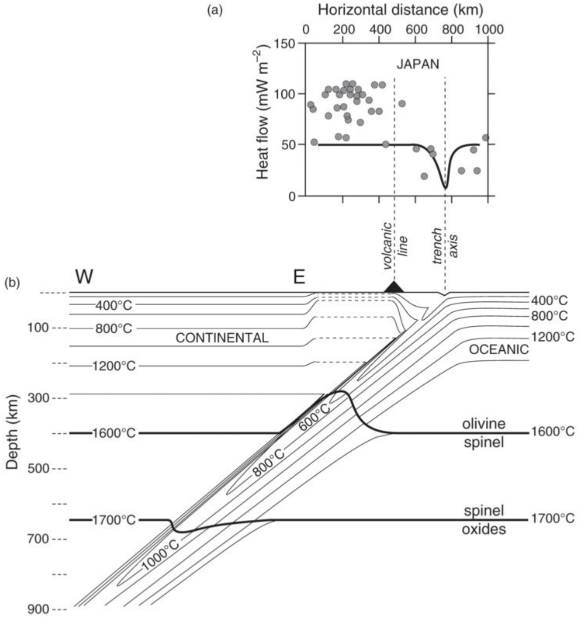


Fig. 9.23 (a) Comparison of heat-flow measurements across the Japanese trench with the theoretical heat flow (solid curve). Source: M.N. Toksöz, J.W. Minear and B.R. Julian, Temperature field and geophysical effects of a downgoing slab, *J. Geophys. Res.* 76, 1113–1138, 1971. (b) The thermal structure of a subduction zone and back-arc region showing the possible isotherms in the cold subducting plate and the thermal effects of the olivine–spinel and spinel–oxide phase changes.

Source: G. Schubert, D.A. Yuen and D.L. Turcotte, Role of phase transitions in a dynamic mantle, *Geophys. J. R. Astr. Soc.* 42, 705–735, 1975.

Heat conducted from the mantle is not the only heat source that must be taken into account in modeling the thermal structure of the subducting slab. An important additional source is the frictional heating that results from shear deformation at the surfaces of the slab where it is in contact with the mantle. In the upper part of a subduction zone the shear heating melts the basaltic layer of the oceanic lithosphere and forms a layer of eclogite in the top of the slab. The high density of the eclogite causes a positive gravity anomaly (see

Fig. 4.31), and adds to the forces propelling the slab downward. The phase transition in which the open structure of olivine-type minerals converts to a denser spinel-type structure normally takes place at a depth of 410 km. The phase transition depends on temperature and pressure. Laboratory experiments indicate that it takes place at lower pressure at low temperature than at high temperature. Consequently it occurs at shallower depths within the cold plate than in the adjacent mantle. As a result, the transition depth is deflected upward by about 100 km. The transition is exothermic and the latent heat given out in the transition is an additional heat source that contributes to the thermal structure of the subduction zone. The transition also results in a density increase, which adds to the forces driving the plate downward.

The deeper transition at 660 km is less well understood. High temperature apparently causes it to take place at higher pressure, and so the depth of occurrence is deflected downward inside the subducting slab. It is uncertain if the transition is endothermic, absorbing heat from the environment, or exothermic as assumed in the model in Fig. 9.23. An endothermic phase change has the effect of reducing the density, and acts against the other downward forces on the slab.

Although other, slightly different models have been derived for the temperature distribution in the descending slab, they all have in common the downward deflection of isotherms in the cold descending slab. The heat flow can be computed for a given thermal model. When compared with the observed heat flow on a profile across the subduction zone, the models fail to explain adequately the high heat flow observed on the overriding plate (Fig. 9.23). Volcanic activity is partly responsible, fed by magmas produced by partial melting of oceanic crustal material in the descending slab and of the upper mantle in the overriding plate. Shallow melting is promoted by water from the subducting plate and generates basaltic magma; deeper melting involves less water and results in andesitic magma. When the overriding plate is continental, volcanic chains form along the continental margin parallel to the deep oceanic trench. The volcanicity is typified by the eruption of both basaltic and andesitic lavas. The lavas are more felsic than those formed when two oceanic plates collide, which may imply that they include melted material from the upper mantle of the overriding continental plate.

When two oceanic plates converge, a volcanic arc is formed on the overriding plate. Behind the arc, high heat flow on the overriding plate is related to back-arc spreading, in which new oceanic crust is generated by the intrusion of basaltic magma from partial melting in the upper mantle. This form of sea-floor spreading produces a marginal basin behind the island arc. The intrusion of magma is not confined to a single location, as at a ridge axis, but is spread diffusely in the basin. Consequently, the stripes of lineated oceanic magnetic anomalies characteristic of sea-floor spreading at ridge systems are missing or at best weakly defined in a marginal basin.

9.9 Mantle Convection

It is well accepted that thermally driven convection takes place in the mantle and that it is probably the most important mechanism in geodynamic processes. There are several reasons for these conclusions. The evidence summarized in Section 5.5 demonstrates that the mantle has a viscoelastic rheology. The passage of seismic compressional and shear waves through the mantle attest that it reacts as a solid to abrupt stress changes. Yet, observations of post-glacial isostatic uplift and long-term movements of the rotation axis indicate that the mantle is capable of viscous flow when stressed over long time intervals. The surmised temperature distribution in the mantle implies that, although conduction is mainly responsible for heat transfer in the lithosphere, convection is the predominant process deeper in the mantle, involving mass transfer by sub-solidus creep. Applying the theory of thermal convection to the mantle and using the best available estimates of physical parameters indicates that robust convection must be taking place.

9.9.1 Thermal Convection

The conditions for convection to occur (see Section 9.4.2) reflect a balance between causal forces due to thermal expansion and resistive effects due to viscosity and thermal diffusivity. When a fluid is heated, thermal expansion gives rise to an upward buoyancy force. This produces instability, which is partly counteracted by diffusion of heat into the surrounding fluid by thermal conduction. As soon as a volume of the fluid starts to rise in response to the buoyancy force, its motion is resisted by viscous forces. The effects are familiar to anyone who has heated a pan of thick soup or porridge. If the pan is heated too rapidly, or the instructions to "stir constantly" are ignored, the soup may stick to the bottom of the pan and become charred. This happens because the viscosity of the fluid is initially too large to allow convection. Despite the large temperature gradient between the hot bottom of the pan and the cool surface of the liquid, conduction is unable to transport heat away from the bottom of the pan fast enough to avoid charring. When heat flow by conduction reaches a critical limit, convection can begin.

The onset of convection in a fluid layer heated from beneath was first described in 1900 by H. Bénard on the basis of laboratory experiments. He noted that a hexagonal pattern of cells forms on the surface of the layer (Fig. 9.24). Hot fluid rises to the surface in the middle of each cell; at the surface it spreads out and cools. Adjoining cells come in contact at narrow margins, where the cooled fluid sinks back into the layer. Each cell has a rectangular cross-section in the vertical plane. A satisfactory theory of Bénard's observations was derived in 1916 by Lord Rayleigh. Although it applies to an ideal scenario (a horizontal layer with stress-free upper and lower boundaries, heated from below, and with a constant temperature on the upper surface) the theory permits approximate estimates for more complex convection in the spherical Earth.

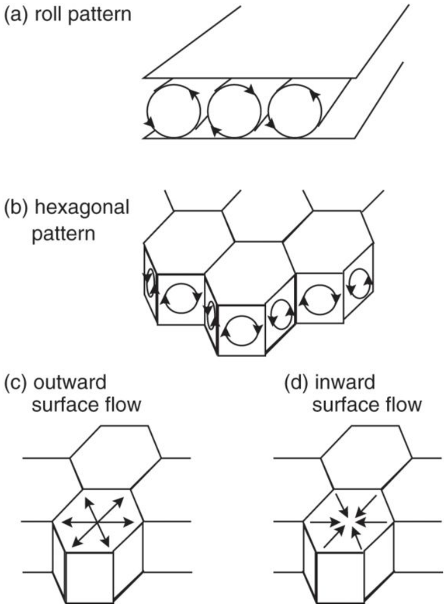


Fig. 9.24 Some patterns of steady convection in a plane layer heated from below. (a) Convection rolls, (b) vertical flow in hexagonal patterns, for which the surface flow may be (c) outward away from or (d) inward toward the center of the cel.

Source: F.H. Busse, Fundamentals of thermal convection, in: *Mantle Convection: Plate Tectonics and Global Dynamics*, W.R. Peltier, ed., pp. 23–95, New York: Gordon and Breach, 1989.

The flow of a viscous fluid is governed by the Navier–Stokes equation, one of the most important equations in geophysics. It describes the conservation of momentum in the fluid, which in its simplest form means balancing several terms that express the driving forces exerted by pressure gradient and buoyancy against the viscous and inertial forces that resist motion. The ratio of viscous diffusion to thermal diffusion is expressed by the dimensionless *Prandtl* number, Pr, defined as

$$\mathrm{Pr} = \frac{\nu}{\kappa}, \tag{9.43}$$

where ν is the kinematic viscosity ($\nu = \eta/\rho$) and κ is the thermal diffusivity ($\kappa = k/(\rho c_p)$). The Prandtl number describes the importance of heat transfer by motion (momentum) of the fluid relative to thermal conduction. A fluid with low Pr flows easily; this is the case in the outer core, where Pr ~ 0.1 (Table 9.6); with increasing Pr the fluid becomes more viscous and thermal conduction loses importance. In the mantle $\nu \approx 10^{17}$ $m^2\ s^{-1}$ and $\kappa \approx 10^{-6}$ $m^2\ s^{-1}$, so that $Pr \approx 10^{23}$. The high Prandtl number means that heat transfer by conduction is unimportant.

Table 9.6 Physical parameters for mantle convection models

Mean values of gravity in the mantle and outer core are 10 m s^{-2} and 8 m s^{-2} respectively. Rayleigh numbers are computed with Eq. (9.18) for a superadiabatic gradient of 1 K km^{-1}. Flow velocities are assumed to be ~3 cm yr^{-1} in the mantle and ~10 km yr^{-1} in the core near its surface.

Some parameters are poorly known; values in the literature vary by up to an order of magnitude.

Physical parameter	Units	Upper mantle (70–660 km)	Whole mantle (70–2890 km)	Outer core (2890–5150 km)
Estimated parameters				
Layer thickness (*D*)	km	590	2820	2260
Expansion coefficient (α)	K^{-1}	2×10^{-5}	1.4×10^{-5}	10^{-5}
Density (ρ)	kg m^{-3}	3700	5000	11,000
Specific heat (c_p)	J kg^{-1} K^{-1}	1250	1250	650
Thermal conductivity (*k*)	W m^{-1} K^{-1}	10	10	100
Dynamic viscosity (η)	Pa s	5×10^{20}	5×10^{21}	10^{-3}
Calculated parameters				

Thermal diffusivity (κ)	$m^2\ s^{-1}$	2.2×10^{-6}	1.6×10^{-6}	1.4×10^{-5}
Kinematic viscosity (ν)	$m^2\ s^{-1}$	10^{17}	10^{18}	10^{-7}
Rayleigh number (Ra_T)	–	8×10^{4}	5.5×10^{6}	1.6×10^{30}
Prandtl number (Pr)	–	6×10^{22}	6×10^{23}	7×10^{-3}
Reynolds number (Re)	–	4×10^{-21}	3×10^{-21}	8×10^{9}
Nusselt number (Nu)	–	4	18	10^{9}

Data sources: Lambeck et al., 2017; Ricard, 2009; Jarvis and Peltier, 1989.

Thermal diffusivity and viscosity act as stabilizing influences in a heated fluid. If heating is slow enough, the temperature gradient adjusts to transfer the heat by conduction, remaining close to the adiabatic gradient. Convection becomes possible when the real temperature gradient exceeds the adiabatic gradient; the difference θ is called the *superadiabatic gradient*. The excess heat expands the fluid, causing the buoyancy force. When this becomes larger than the viscous resistance, convection ensues. The ratio of the competing forces is embodied in the *Rayleigh* number (Ra_T). Equation (9.18), repeated here, gives the Rayleigh number for convection due to a superadiabatic temperature gradient in a fluid layer of thickness D as

$$Ra_T = \frac{g\alpha\theta}{\kappa\nu} D^4, \tag{9.44}$$

where g is gravity and α is the coefficient of thermal expansion.

The superadiabatic gradient is not the only source of power for convection. Although radioactive heat generation in mantle materials is small (see Section 9.5.1), it can still contribute to convection. If Q is the radiogenic heat production in a layer of thickness D, we can invoke Eqs. (9.15) and (9.21) and replace θ in the above equation by (QD/k), where k is the thermal conductivity. This allows us to define a second Rayleigh number (Ra_Q) for convection driven by radiogenic heat:

$$\mathrm{Ra}_Q = \frac{g\alpha Q}{k\kappa\nu} D^5 . \tag{9.45}$$

Convection is initiated when the Rayleigh number exceeds a critical value, Ra_c, which is dependent on the geometry of the flow and the boundary conditions on the upper and lower surfaces. In Rayleigh–Bénard convection the top and bottom of the horizontal layer are stress-free; the critical Rayleigh number is $\mathrm{Ra}_\mathrm{c} = 658$. If the top and bottom of the layer are rigid boundaries at which the horizontal velocity vanishes, $\mathrm{Ra}_\mathrm{c} = 1708$. Computed Rayleigh numbers Ra_T for convection driven by the superadiabatic temperature gradient for viscous flow in the upper mantle, whole mantle, and core greatly exceed Ra_c (Table 9.6). Reasonable estimates of the radiogenic heat produced in the mantle give even larger values for Ra_Q.

9.9.2 Convection at High Rayleigh Numbers

The large Rayleigh numbers for the entire mantle and for the upper layer alone indicate that each region of the sub-lithospheric mantle is capable of convection. The Rayleigh number for whole-mantle convection is so much larger than the critical value $\mathrm{Ra_c}$ that vigorous mantle convection must be expected. This does not imply rapid flow in normal terms. The speed of flow in the mantle is usually assumed to be of the same order as the rate of motion of tectonic plates, about 5–10 cm $\mathrm{yr^{-1}}$ on average. As long as the flow rate V_c is low, adjacent lamina of the fluid move past each other under the conditions for Newtonian viscosity (Section 5.2.2). At faster flow rates this condition breaks down, and the flow becomes turbulent. The conditions favoring turbulence are high momentum (ρV_c) and large scale (D) of the flow, whereas it is inhibited by high viscosity η. These factors are contained in the Reynolds number, Re, defined as

$$\mathrm{Re} = \frac{\rho V_c D}{\eta}. \tag{9.46}$$

Turbulent flow sets in when Re exceeds a critical value; this depends on the geometry of the flow, but in many situations is around 10^3–10^5. Representative values for mantle parameters (Table 9.6) and an assumed velocity comparable to that of plate tectonic motions ($V_c \approx 30\ \mathrm{mm\ yr^{-1}} = 10^{-9}\ \mathrm{m\ s^{-1}}$) give a Reynolds number around $\mathrm{Re} = 3 \times 10^{-21}$, which is so small that turbulence is negligible. Similar results are found by considering the upper or lower mantle alone. Clearly, although mantle convection involves high Rayleigh numbers (implying vigorous convection on a geological timescale), it takes place by laminar flow.

The effect of convection is to replace conduction as the principal mechanism of heat transfer. A measure of the relative effectiveness of the two processes of heat transfer is the Nusselt number, Nu. This is defined as the ratio of the heat transport with convection to the heat transport without convection. In the absence of radiogenic heat sources, the heat transport with convection is determined by the Rayleigh number Ra_T, while the non-convective heat transport is expressed by the critical Rayleigh number Ra_c. The Nusselt number depends on the ratio of these two numbers and can be written as

$$\mathrm{Nu} = \beta\left(\frac{\mathrm{Ra}_T}{\mathrm{Ra}_c}\right)^S, \tag{9.47}$$

where the coefficient β and the exponent S are functions of the aspect ratio of the convection cells. Mathematical evaluation of the problem of Rayleigh–Bénard convection with stress-free upper and lower boundaries gives $\beta \approx 1$ and $S = 1/3$, and, since in this case $\mathrm{Ra}_c \approx 10^3$, the Nusselt number has the simpler form

$$\mathrm{Nu} \approx 0.1(\mathrm{Ra}_T)^{1/3}. \tag{9.48}$$

Using the estimated values of Ra_T in Table 9.6 gives Nu ~ 4 for layered convection in the upper mantle and Nu ~ 18 for whole-mantle convection. A value $\mathrm{Nu} > 1$ implies that heat transfer by convection is dominant, which is evidently the case throughout the mantle.

Once convection has been initiated, the boundary conditions determine the shapes of the convection cells. In Rayleigh–Bénard convection the aspect ratio of a cell – the ratio of its horizontal dimension to its vertical one – is $2^{1/2} = 1.41$; when the layer has rigid boundaries the cell aspect ratio is 1.01. Hence, the horizontal extent of a convection cell is comparable with the layer thickness. This has implications for convection in the mantle. If we assume that the scale of mantle convection is represented by the pattern of the plate boundaries (Fig. 2.7), we can estimate the horizontal dimensions of the convection cells. It is evident that they must have very different sizes. The ridge-to-trench horizontal distances across major plates are in the range 2000–10,000 km, with an average of about 5000 km. This is larger than the maximum thickness of the convecting layer, whether we assume convection to be restricted to the upper mantle or to occupy the entire 2900 km thickness of the sub-lithospheric mantle. Thus, if convection is uniform through the whole mantle, the aspect ratios of at least some cells must be much larger than unity (Table 9.7). The reason for this is the rigidity of the cold upper boundary formed by the lithosphere, which inhibits the breakup of the fluid flow into cells with smaller horizontal extents.

Table 9.7 Approximate aspect ratios of some mantle convection cells, estimated from the horizontal dimensions of the overlying lithospheric plates

Plate	Upper mantle convection	Whole mantle convection
Pacific	14	3.3
North American	11	2.6
South American	11	2.6
Indian	8	2.1
Nazca	6	1.6

Source: after Turcotte and Schubert, 1982, table 7.5.

9.9.3 Models of Mantle Convection

Thermal parameters for the mantle favor convection as the most important mechanism for transferring heat on a geological timescale. There is some uncertainty as to the form it takes. The seismic discontinuities at 410 km and 660 km depth, which bound the upper mantle transition zone, represent mineral phase changes rather than compositional differences, such as define the crust–mantle and core–mantle boundaries. The upper discontinuity marks the olivine–spinel phase change, the lower one represents the phase change from spinel to a perovskite structure (Section 7.4.4.1), specifically from the mineral ringwoodite to bridgmanite, which exemplifies the lower mantle. Each phase change is accompanied by changes in density and elastic parameters. Because there is no compositional change, mass can be carried by convection currents across these discontinuities. An important increase in viscosity occurs around the transition zone (Fig. 9.25), by an order of magnitude from the upper mantle to the lower mantle (Table 9.6). The 660 km discontinuity is close to the maximum depth of seismicity in subduction zones. Subducting tectonic plates may be held up at this discontinuity, flattening their downward trajectory. However, seismic tomography has demonstrated that the descending slabs can penetrate the discontinuity and sink deep into the lower mantle. Two contrasting models have been proposed for the mechanism by which mantle convection takes place.

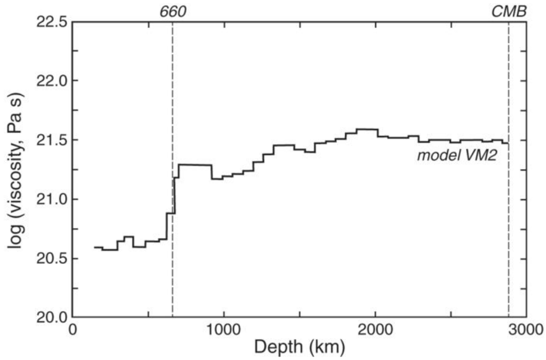


Fig. 9.25 Model profile of viscosity η in the mantle, derived from analysis of post-glacial isostatic adjustment in Fennoscandia and the Canadian shield, taking into account accompanying changes in rotational parameters (based on Peltier, 2004).

The *layered convection* model comprises separate convection cells in the upper and lower mantle, with the 660 km discontinuity acting as the boundary that separates them. The upper and lower mantle are well mixed individually, but the separation of the flow patterns at the 660 km discontinuity implies different chemical compositions, which excludes the transfer of mass across the discontinuity. This two-layer model was proposed for compatibility with geochemical differences among oceanic basalts. Mid-ocean ridge basalts (or MORB) have a more depleted trace element geochemistry than oceanic island basalts (or OIB), which occur on oceanic plates well away from plate margins and which are typically associated with hotspots (Section 2.8, Fig. 2.21). The OIB have anomalously high $^3He/^4He$ isotopic ratios. This is interpreted as evidence that the OIB originate from a source that is older and geochemically more enriched than the MORB. Their compositions are varied and this suggests that they derive from various sources that have radiometric ages as old as 1 Gyr (i.e., older than the MORB but younger than the age of the Earth).

Layered convection therefore has consequences for thermal conditions in the deep mantle and core. In the absence of convective flow across the 660 km level, heat can only cross it by conduction. The discontinuity would therefore act as a thermal boundary, with a large temperature change of perhaps 500–1000 K across it. Thus, the temperature profile in the lower mantle, although maintained adiabatic by the convection, would be 500–1000 K higher than in whole-mantle convection. A hotter lower mantle results in a smaller temperature change across the core–mantle boundary. This would result in a less-steep temperature gradient in the D″ layer, and a lower heat flux from the core. The long-term cooling rate of the core would thereby be reduced.

The *whole-mantle convection* model entails a net flow of matter across the 660 km discontinuity (Fig. 9.26). Convection thus ensures that the entire mantle is well mixed mechanically, and the phase changes at 410 km and 660 km have only a small effect on the temperature gradient. The mantle therefore lies between two thermal boundary layers. It is heated from below by heat that flows across the core–mantle boundary, which constitutes a hot, lower thermal boundary. It loses heat externally by conduction through a cold upper thermal boundary, the lithosphere, across which the temperature drops by more than 1000 K. These thermal boundaries influence the pattern of convective flow in the intervening mantle. Deep in the mantle, temperatures are high and the mantle flows by thermally activated diffusion creep (Section 5.2.3.2). At shallower depths in the mantle the deformation mechanism is dislocation creep, which is also a thermally activated process. The mantle is solid for the passage of seismic waves, but thermally activated deformation allows it to flow over long geological timescales and thereby makes convection the most important thermal transport mechanism.

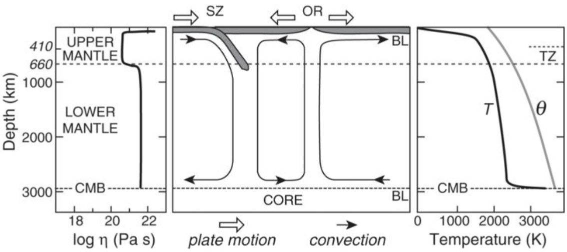


Fig. 9.26 Hypothetical profiles of viscosity η (*left*), temperature *T* and solidus temperature *θ* (*right*), and a whole-mantle convection flow pattern (*center*). Abbreviations: SZ, subduction zone; OR, oceanic ridge; TZ, upper mantle transition zone; BL, boundary layers; CMB, core–mantle boundary.

Source: W.R. Peltier, G.T. Jarvis, A.M. Forte and L.P. Solheim, The radial structure of the mantle general circulation, in: *Mantle Convection: Plate Tectonics and Global Dynamics*, W.R. Peltier, ed., pp. 765–815, New York: Gordon and Breach, 1989.

An order-of-magnitude estimate of times involved in whole-mantle convection can be made by comparison with the rates of tectonic plate motions. Lithospheric plates are driven mainly by slab pull and trench suction at subduction zones. Drag forces caused by convection in the underlying mantle are insignificant (see Fig. 2.27). A convection current beneath the plate must therefore be moving at a speed similar to that of the plate. Rates of separation and closure between the plates are around 10–100 mm yr^{-1}, with median ~50 mm yr^{-1} (Fig. 2.7). An individual plate moves at half this rate, so convection may take place at ~25 mm yr^{-1} below the lithosphere. It may be much slower in the more viscous deep mantle. At this rate it would take ~100 Myr for mantle material to traverse the ~2500 km from the core–mantle boundary to the lithosphere, assuming no hindrance at the 660 km transition. For a Rayleigh–Bénard convection cell with ~2:1 aspect ratio (Table 9.7), this implies a whole-mantle overturn time on the order of 600 Myr or longer.

The whole-mantle convection model is successful in explaining the size distribution of lithospheric plates. Seismological data (e.g., earthquake first-motion studies) accumulated and led to an increase in the number of known and presumed plates. Additionally, space geodesy – in particular GPS measurements – made it possible to measure strain rates in tectonically active zones. As a result the number of presumed tectonic plates has increased to 52 (and many more very small plates have been proposed). When the 52 best-known plates are ranked in order of size, from largest to smallest, they are distributed in three groups. Ranking by size is equivalent to counting the number of plates larger than a given size. A log–log plot of a plate’s rank against its area shows linear segments (Fig. 9.27), implying power-law relationships. The main group containing most of the plates therefore fulfills the requirements of a fractal size distribution. This implies that the plate sizes are self-organizing, i.e., their sizes result from interactions between each other. Geodynamic modeling of the formation of plates in a scenario of whole-mantle convection successfully explains the observed size distribution. The model treats the lithosphere as a rigid upper boundary to mantle convection, which fragments into a few large plates and many small plates due to the bending-stresses at subduction zones. The strength of the lithosphere is the deciding factor in determining the distribution of plate sizes produced in the fragmentation.

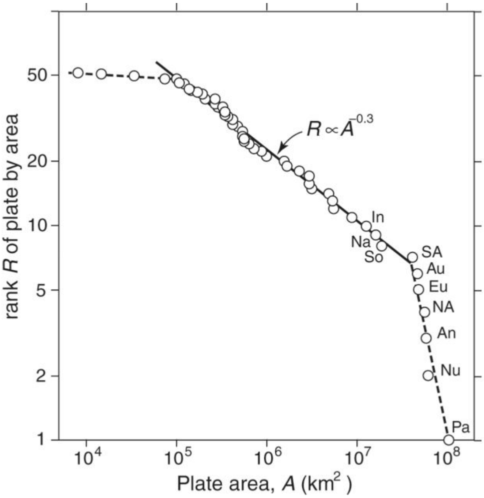


Fig. 9.27 Size distribution of lithospheric plates, ranked in order of increasing area (data source: Bird, 2003). Letters identify the largest plates: Pa, Pacific; Nu, Nubia; An, Antarctica; NA, North America; Eu, Eurasia; Au, Australia; SA, South America; So, Somalia; Na, Nazca; In, India.

9.9.4 The D″ Layer and Superplumes

The seismic D″ layer at the base of the mantle (Section 7.4.4.2), between about 2700 km depth and the hot boundary layer at the core–mantle boundary, has a complex seismic velocity structure. It evidently has different physical properties than the mantle above it, and has an important role in mantle dynamics. A phase transition takes place around this depth, in which perovskite, the magnesium silicate that typifies the lower mantle, is thought to convert to a high-pressure phase called post-perovskite. This mineral is stable at the pressure (120 GPa) and temperature (2500 K) in the D″ layer. The heat flux from the core to the mantle diminishes the rigidity of the layer and lowers seismic velocities in it. The layer is thin but hot, and its viscosity is presumed to be much lower than that of the overlying mantle. The topography of the core–mantle boundary has been explored by seismic waves reflected from the core or passing it at grazing incidence. The thickness of the D″ layer appears to be uneven and it is seismically heterogeneous, with anomalous velocity gradients. Seismic shear waves are polarized and anisotropic in this layer, which results in shear-wave splitting. The velocities of horizontally polarized S-waves are slightly faster than vertically polarized S-waves. This may be due to preferred orientation of post-perovskite minerals.

The lower mantle contains other seismically anomalous regions. Seismic tomography has identified two large regions at the base of the mantle – one beneath the Pacific Ocean, the other under Africa – in which seismic P-wave and S-wave velocities are lowered. These large low-velocity provinces (LLVP) are continental in size (the Pacific anomaly is 3000 km across) and extend about 1000 km above the core–mantle boundary. They have been called *superplumes* or *thermochemical piles*, with implications for their interpreted geodynamic or geochemical function. The concept of a superplume is that of a region that is hotter than the surrounding lower mantle, making it thermally buoyant and resulting in mantle upwelling. The distribution of hotspots on the lithosphere is concentrated above the margins of the superplumes, but a direct link has not been established. An alternative interpretation of the low shear-wave velocities in LLVP is that they result from chemical differences with the lower mantle. Geochemical arguments suggest that they may represent ancient subducted lithosphere that has accumulated at the core–mantle boundary, or they may consist of primordial material that remains from the early differentiation of the Earth.

9.9.5 Mantle Plumes and Hotspots

Although the nature and origin of superplumes and their meaning for mantle dynamics are still uncertain, the existence of these features is not in doubt. Seismic tomography and shear-wave studies have established their existence and many of their characteristics. More mysterious are *mantle plumes*. These predominantly oceanic features were originally envisaged by W. J. Morgan in 1971 as narrow conduits of hot material rising from the deep mantle to feed surface hotspots. A precursor to this concept was suggested by J. T. Wilson in 1963, when plate tectonics and mantle dynamics were not yet understood, to explain the origin of the Hawaiian islands. Plumes have been successfully modeled in computer and laboratory experiments and have become intrinsic to models of mantle dynamics. A postulated source for the low-viscosity hot magma in the plumes is the D″ layer.

A mantle plume is a relatively narrow feature, some 100–500 km in cross-section, that facilitates the upwelling of low-viscosity hot rock through the more viscous mantle. A young plume is visualized as rising through the mantle behind a broader plume head, forming a mushroom-shaped structure (Fig. 9.28). Some plumes may not breach the lithosphere, but instead melt to form a basaltic magma as a result of decompression. Other plumes may penetrate the entire mantle and reach the surface, where they are evident as places of persistent volcanism, high regional and geoid topography, and local high heat flow. Such regions, known as *hotspots* (see also Section 2.8) are found in the oceans and on the continents, within plates and on plate margins. The plumes that feed them are thought to remain nearly fixed in position for long periods of time, and so the hotspots move only very slowly relative to the deep mantle. They therefore provide a comparatively stationary network of positions that can be used as a reference for studies of plate tectonic motions. By assuming that they are fixed, paleomagnetic pole positions can be referred to a hotspot framework and used to calculate *true polar wander*, the motion of the rotation axis relative to the mantle over millions of years (Section 12.2.4.5).

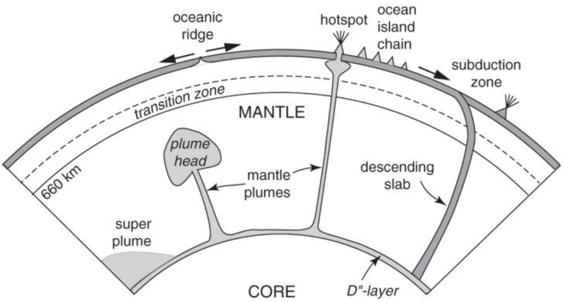

oceanic ridge
hotspot
ocean island chain
subduction zone
transition zone
MANTLE
660 km
plume head
mantle plumes
descending slab
super plume
CORE
D″-layer

Fig. 9.28 Cross-section through the mantle, showing suggested relationships between plate tectonic elements, mantle plumes, and the D″ layer.

Source: F.D. Stacey, *Physics of the Earth*, 513 pp., Brisbane, Australia: Brookfield Press, 1992.

A mantle plume is depicted as relatively thin and composed of material that is hot and has a lower viscosity than the surrounding mantle. Shear waves in such a plume are expected to have lower than average P-wave and S-wave velocities, which can indeed be imaged in the upper mantle beneath most hotspots using seismic tomography. However, an objection to the plume concept is that continuous plumes emerging from the lower mantle to the surface have not been observed unequivocally in seismic tomography studies. This may be a consequence of insufficient seismic data coverage, especially at greater depths below the 410 km and 660 km discontinuities. Seismologists point out that the petrological differences between MORB and OIB could result from the OIB being formed in local zones of melting related to extension of the lithosphere. The models of mantle dynamics are very convincing, but the objections to them are real and as yet unresolved.

9.10 *Suggestions for Further Reading*

Introductory Level

Davies, G. F. 2011. *Mantle Convection for Geologists*. Cambridge: Cambridge University Press.

Intermediate and Advanced Level

Davies, G. F. 1999. *Dynamic Earth: Plates, Plumes and Mantle Convection*. Cambridge: Cambridge University Press.

Foulger, G. R. 2010. *Plates vs Plumes: A Geological Controversy*. New York: Wiley-Blackwell.

Jessop, A. M. 1990. *Thermal Geophysics*. Amsterdam: Elsevier.

Peltier, W. R. (ed.) 1989. *Mantle Convection: Plate Tectonics and Global Dynamics*. New York: Gordon and Breach.

Schubert, G., Turcotte, D. L., and Olson, P. 2001. *Mantle Convection in the Earth and Planets*. Cambridge: Cambridge University Press.

Stacey, F. D. and Davis, P. M. 2008. *Physics of the Earth* (4th ed.). Cambridge: Cambridge University Press.

Turcotte, D. L. and Schubert, G. 2014. *Geodynamics* (3rd ed.). Cambridge: Cambridge University Press.

9.11 Review Questions

1. What is meant by temperature? What is meant by heat?

2. What are the processes by which heat can be transferred? What is the relative importance of each process in (a) the crust, (b) the mantle, (c) the outer core, and (d) the inner core?

3. Sketch how (a) temperature and (b) the melting point (solidus) vary with depth in the Earth's interior.

4. How is heat flow defined? How is it measured (a) on the continents and (b) in the oceans?

5. What factors determine the depth of penetration of solar energy into the Earth? What precautions does this impose for measuring heat flow?

6. Why is the average oceanic heat flow higher than the average continental heat flow?

7. How does heat flow vary with distance from an oceanic ridge?

8. Which regions of the Earth have (a) the highest and (b) the lowest heat flow?

9. Discuss the statement: "The internal heat of the Earth causes the formation of mountains and the external heat of the Sun causes their destruction."

9.12 Exercises

1. List and compare the various factors that may influence the measured temperature gradient at a depth of 5 m in (a) a deep drillhole in oceanic sediments and (b) a continental well that encounters the groundwater table at 2 m depth.

2. A shallow circular pond 100 m in diameter freezes solid during a very cold night. The pond is in a geothermal area in which the temperature reaches 40°C at 200 m depth. The thermal conductivity of the intervening rock is 3.75 W m^{-1} K^{-1} and the latent heat of fusion of ice is 334 kJ kg^{-1}. Neglecting other heat sources, calculate the mass of ice that melts per hour due to the geothermal gradient.

3. Assuming a constant geothermal gradient of 30°C per kilometer, estimate what percentage of the Earth's volume is hotter than the temperature of molten lava at atmospheric pressure. Why is the deeper interior of the Earth not entirely molten?

4. The mean global heat flow at the Earth's surface is 87 mW m^{-2}. Calculate the time in years needed for the mantle and core to cool by 100°C, with the following assumptions: (1) the Earth's mantle and core cool as a homogeneous unit; (2) 20% of the observed heat flow at the Earth's surface comes from below the lithosphere; (3) the lithospheric thickness is 100 km; (4) thermal effects from the lithosphere itself may be ignored. Relevant properties of the mantle and core are: mean density 5650 kg m^{-3}, specific heat 400 J kg^{-1} K^{-1}.

5. A temperature gradient of 35°C km^{-1} is measured in the upper few meters of sediments covering the ocean floor. If the mean thermal conductivity of oceanic sediments is 1.7 W m^{-1} K^{-1}, calculate the local heat flow. How far do you think the sampling site is from the nearest active ridge?

6. What heat-flow values would you expect at the locations of the oceanic magnetic anomalies with numbers C5N, C10N, C21N, C32N, M0? Interpret the ages of the anomalies from Fig. 12.32 and use the heat-flow model GDH1 (Eq. 9.40) for the cooling of oceanic lithosphere.

7. Using the relationships in Eq. (9.41), estimate the approximate depths of the ocean at the locations in Exercise 6.

8. Assuming that the Earth initially had a uniform temperature throughout and has been cooling by conduction only, use the solution for the one-dimensional cooling of a semi-infinite half-space (Eq. 9.34 and Box 9.2) to derive Eq. (9.2) for Kelvin's estimated age of the Earth.

9. The temperature in the near-surface layers of the Earth's crust varies cyclically with daily, annual, and longer periods. For a surface temperature variation given by $T = T_0 \cos\omega t$, the temperature variation at depth z and time t is described by:

$$T(z,t) = T_0 \exp\left(-\frac{z}{d}\right) \cos\left(\omega t - \frac{z}{d}\right)$$

$$d = \sqrt{\frac{2\kappa}{\omega}}; \qquad \kappa = \frac{k}{\rho c_p}; \qquad \omega = \frac{2\pi}{\tau},$$

where τ is the period of the variation, k is the thermal conductivity, c_p is the specific heat, and ρ is the density. For surface sediments assume $k = 2.5$ W m K^{-1}, $c_p = 1000$ J kg^{-1} K^{-1}, and $\rho = 2300$ kg m^{-3}.

(a) Calculate the phase difference (in days) between the temperature variation at the surface and at depths of 2 m and 5 m, respectively. Perform the calculations for both the daily and annual temperature fluctuations.

(b) Assuming that the range in surface temperatures between summer and winter is 40°C, calculate the depth at which the annual temperature range is 5°C. How large (in weeks and days) is the phase difference between the surface temperature and the actual temperature at this depth?

10. The average daily temperature in northern Canada is +10°C in July and –20°C in January. Using the heat conduction equation, calculate the thickness of the permafrost layer (below which the ground is permanently frozen). Relevant physical properties of the ground are: thermal conductivity $k = 3$ W m K^{-1}, specific heat $c_p = 840$ J kg^{-1} K^{-1}, density $\rho = 2700$ kg m^{-3}.

11. The half-spreading rate at an oceanic ridge in the middle of a symmetric ocean basin bounded by subduction zones is 44 mm yr^{-1}. The ridge is 1000 km long and the distance from the ridge to each subduction zone is 2000 km. If the oceanic heat flow varies with crustal age, as in Eq. (9.40), calculate how much heat is lost per year from the ocean basin.

9.13 Computational Exercises

Jupyter notebook computational exercises can be found online for this chapter at www.cambridge.org/FoG3ed (Heat flow and age).

10

Geoelectricity

Preview

The rocks and minerals that make up the Earth’s crust and mantle have variable electrical properties. This characteristic is made use of in important methods of surveying, to research the structure and properties of the interior. Natural and induced electrical currents conform to the underground pattern of electrical resistivity. We describe how this can be used to locate anomalous regions with mineral enrichment or that have environmental importance. Electromagnetic methods, employed on the ground and from the air, are particularly important in mineral exploration. Ground penetrating radar has become an important tool of environmental and archeological research. These methods are limited to the top several hundred meters of the crust, but magnetotelluric methods can probe deeper into the lithosphere and mantle. The high-frequency variations of the magnetic field induce currents in the solid Earth and in the oceans, delivering valuable information about the electrical conductivity in the upper mantle.

10.1 Introduction

Electric charge is a fundamental property of nature. The name *electric* derives from the Greek word for amber ("elektron"), the naturally occurring fossilized resin of coniferous trees that has been used since antiquity in the making of jewelry. The Greek philosopher Thales of Miletus (*c.* 600 BC) is credited with first reporting the power of amber, when rubbed with a cloth, to attract light objects. The ancient sages could not understand this behavior in terms of their everyday world, and so, together with the power of *magnetism* possessed by natural lodestone (see Section 11.1.1), electricity remained a wonderful but unknown phenomenon for more than two millennia. In AD 1600 the English physician William Gilbert summarized previous investigations and extant knowledge in the first systematic study of these phenomena.

In the following century it was established that there were two types of electric charge, now referred to as positive and negative. Objects that carried like types of charge were observed to repel each other, and those that carried opposite types were attracted to each other. In 1752 the American statesman, diplomat, and scientist Benjamin Franklin performed a celebrated experiment; by flying a kite during a thunderstorm, he established that lightning is an electrical phenomenon. Having survived this risky endeavor, Franklin developed the far-sighted theory that electricity consisted of an omnipresent fluid, and that the different types of charge represented surplus and scarcity of this fluid. This view strikingly resembles modern theory, in which the "fluid" consists of electrons.

The laws of electrostatic attraction and repulsion were established in 1785 as a result of careful experiments by a French scientist, Charles Augustin de Coulomb (1736–1806), who also established the laws of magnetostatic force (Section 11.1.3). Coulomb invented a sensitive torsion balance, with which he could measure accurately the force between electrically charged spheres. His results represent the culmination of knowledge of electrostatic phenomena.

The eighteenth-century concept of electricity as a fluid finds further expression in electrical nomenclature. Electricity is said to flow between charged objects when they are brought into contact, and the rate of flow is called an electric current. The study of the properties and effects of electric currents became possible around 1800, when an Italian physicist, Alessandro Volta (later elevated by Napoleon to the rank of count), invented a primitive electric battery, called a voltaic pile, in which electricity was produced by chemical action. The relationship between the electric current in a conductor and the voltage of the battery was established in 1827 by Georg Ohm, a German physicist. The magnetic effects produced by electric currents were established in the early nineteenth-century by Oersted, Ampère, Faraday, and Lenz. Their contributions are discussed in more detail in a later section (Section 10.6) on the physical origins of magnetism.

10.2 Electrical Principles

Coulomb established that the force of attraction or repulsion between two charged spheres was proportional to the product of the individual electric charges and inversely proportional to the square of the distance between the centers of the spheres. His law can be written as the following equation:

$$F = K\frac{Q_1 Q_2}{r^2},$$

(10.1)

where Q_1 and Q_2 are the electric charges, r is their separation, and K is a proportionality constant. This inverse-square law strongly resembles the law of universal gravitation (Eq. 3.2), formulated by Newton more than a century before Coulomb's law. However, in gravitation the force is always attractive, whereas in electricity it may be attractive or repulsive, depending on the nature of the charges. In the law of gravitation the units of mass, distance, and force are already defined, so that the gravitational constant is predetermined; only its numerical value needed to be measured. In Coulomb's law, F and r are defined from mechanics (as the newton and meter, respectively), but the units of Q and K are undefined. The value of K was originally set equal to unity, thereby defining the unit of electric charge. This definition led to unfortunate complications when the magnetic effects of electric currents were analyzed. The alternative is to define independently the unit of charge, thereby fixing the meaning of the constant K.

The unit of charge is the coulomb (C), defined as the amount of charge that passes a point in an electrical circuit when an electric current of one ampère (A) flows for one second (i.e., 1 C = 1 A s). In turn, the ampère is defined from the magnetic effects of a current (see

Box 11.1). When a current flows in the same direction through two parallel long, straight conductors, magnetic fields are produced around the conductors, which cause them to attract each other. If the current flows through the conductors in opposite directions they repel each other. The ampère is defined as the current that produces a force of 2×10^{-7} N per meter of length between infinitely long, thin conductors that are one meter apart in vacuum. Thus, the unit of charge is defined precisely, if rather indirectly. In the Système Internationale (SI) units, *K* is written as $(4\pi\, \varepsilon_0)^{-1}$, so that Coulomb's law becomes

$$F = \frac{1}{4\pi\varepsilon_0}\frac{Q_1 Q_2}{r^2}, \tag{10.2}$$

where the constant ε_0 is called the *permittivity constant*, or alternatively the *electric constant*. It is equal to $8.854187817 \times 10^{-12}$ $C^2\ N^{-1}\ m^{-2}$.

Modern electrical theory descends from the discovery in 1897 by the English physicist Joseph J. Thomson of the *electron* as the basic elementary unit of electric charge. It has a *negative charge* of 1.602×10^{-19} C. A proton in the nucleus of an atom has an equal *positive charge*. Normally an atom contains as many electrons as it has protons in its nucleus and is electrically neutral. If an atom or molecule loses one or more electrons, it has a net positive charge and is called a *positive ion*; similarly, a *negative ion* is an atom or molecule with a surplus of electrons.

In metals, some electrons are only loosely bound to the atoms. They can move with relative ease through the material, which is called an electrical *conductor*. Metals like copper and silver are good conductors. In other materials, called *insulators*, the electrons are tightly bound to the atoms. Glass, rubber, and dry wood are typical insulators. A perfect insulator does not allow electrons to move through it, whereas a perfect conductor offers no opposition to the passage of electrons. Real conductors offer different degrees of opposition.

A flow of charge, or *electric current*, results when the free electrons in a conductor move in a common direction. A current of one ampère corresponds to a flow of about 6.242×10^{18} electrons per second past any point of a circuit! The direction of an electric current is defined to be the direction of flow of positive charge, which is opposite to the direction of motion of the electrons.

10.2.1 Electric Field and Potential

The force exerted on a unit electric charge by another charge Q is called the *electric field* of the charge Q. Thus, if we let $Q_1 = Q$ and $Q_2 = 1$ in Eq. (10.2), we obtain the equation for the electric field E at distance r from a charge Q:

$$E = \frac{Q}{4\pi\varepsilon_0 r^2}. \tag{10.3}$$

According to this definition, E has the dimensions of newton/coulomb (N C^{-1}).

The term "field" also has another connotation, introduced by Michael Faraday (1791–1867) to refer to the geometry of the lines of force near a charge. Around a positive point charge the field lines are radially outward, describing the (divergent) direction along which a free positive charge would move (Fig. 10.1a); around a negative point charge they are radially inward (convergent) (Fig. 10.1b). The field lines of a pair of opposite point charges diverge from the positive charge, spread apart, and converge on the negative charge (Fig. 10.1c); they give the appearance of drawing the opposite charges together. The combined field of two positive point charges is characterized by field lines that leave each charge and diverge in the space between (Fig. 10.1d); the field lines appear visibly to push the like charges apart. The direction of the electric field at any point is tangential to the electric field line. The strength of the field is represented by the spatial concentration of the field lines. Close to either electrical charge the field is strong and it weakens with increasing distance from the charge. Consequently, work is required to move a charged particle from one point in the field to another. This work contributes to the potential energy of the system.

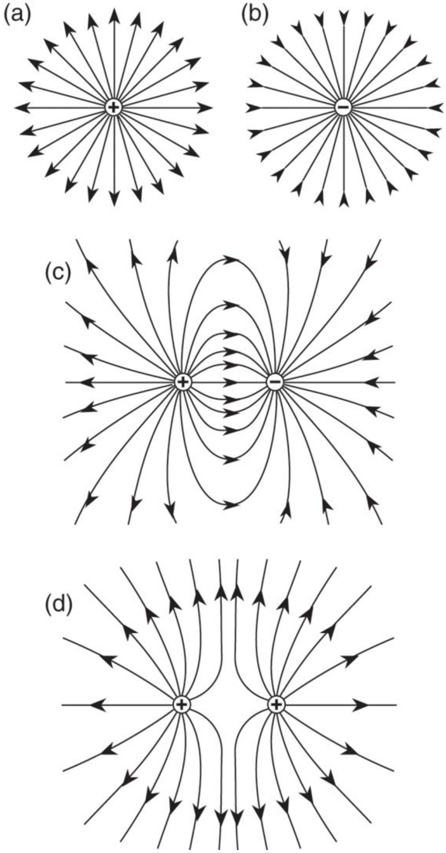


Fig. 10.1 Planar cross-sections of electric field lines around point charges: (a) single positive, (b) single negative, (c) two equal and opposite, and (d) two equal positive charges.

For example, at an infinite distance from a positive charge Q, the repulsive force on a unit positive charge is zero, but at a distance r it is given by Eq. (10.3). The potential energy of the unit charge at r is called the *electric potential* at r; we will denote it U. The units of U are energy per unit charge, i.e., joules/coulomb. If we move a distance dr against the field E, the potential changes by an amount dU equal to the work done against E, which is $(-E\,\mathrm{d}r)$. i.e., $\mathrm{d}U = -E\,\mathrm{d}r$, so that

$$E = -\frac{\mathrm{d}U}{\mathrm{d}r}. \tag{10.4}$$

We can readily compute the electric potential U at r by integration:

$$U = -\int_{\infty}^{r} E \, \mathrm{d}r = -\int_{\infty}^{r} \frac{Q}{4\pi\varepsilon_0 r^2} \mathrm{d}r, \tag{10.5}$$

from which

$$U = \frac{Q}{4\pi\varepsilon_0 r}. \tag{10.6}$$

The energy needed to move a unit charge from one point to another in the electric field of Q is the *potential difference* between the two points. The unit of potential difference is the same as that of U (i.e., joules/coulomb) and is called a *volt*. From Eq. (10.4) we obtain the more common alternative units of volt/meter (V m^{-1}) for the electric field E.

Electric charge flows from a point with higher potential to a point with lower potential. The situation is analogous to the flow of water through a pipe from one level to a lower level. The rate of flow of water through the pipe is determined by the difference in gravitational potential between the two levels. Likewise, the electric current in a circuit depends on the potential difference in the circuit.

10.2.2 Ohm's Law

The German scientist Georg Simon Ohm established in 1827 that the electric current I in a conducting wire is proportional to the potential difference V across it. The linear relationship is expressed by the equation

$$V = IR, \tag{10.7}$$

where R is the *resistance* of the conductor. The unit of resistance is the *ohm* (Ω). The inverse of resistance is called the *conductance* of a circuit; its unit is the reciprocal ohm (Ω^{-1}), which is called a *siemens* (S).

Experimental observations on different wires of the same material showed that a long wire has a larger resistance than a short wire, and a thin wire has a larger resistance than a thick wire. Formulated more precisely, for a given material the resistance is proportional to the length L and inversely proportional to the cross-sectional area A of the conductor (Fig. 10.2). These relationships are expressed in the equation

$$R = \rho \frac{L}{A}. \tag{10.8}$$

The proportionality constant ρ is the *resistivity* of the conductor. It is a physical property of the material of the conductor, which expresses its ability to oppose a flow of charge. The inverse of ρ is called the *conductivity* of the material, denoted σ. The unit of resistivity is the ohm-meter (Ω m). The unit of conductivity is the siemens/meter (S m^{-1}); it may also be given as the reciprocal ohm-meter (Ω^{-1} m^{-1}).

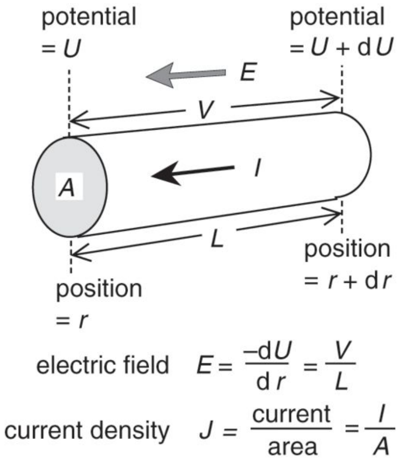

potential
= U
potential
= $U + \mathrm{d}U$
E
V
A
I
L
position
= $r + \mathrm{d}r$
position
= r
electric field $E = \frac{-\mathrm{d}U}{\mathrm{d}r} = \frac{V}{L}$
current density $J = \frac{\text{current}}{\text{area}} = \frac{I}{A}$

Fig. 10.2 Parameters used to define Ohm's law for a straight conductor.

If we substitute Eq. (10.8) for R in Eq. (10.7) and rearrange the terms, we get the following expression:

$$\frac{V}{L}=\rho\frac{I}{A}.$$

(10.9)

The ratio V/L on the left side of this equation is, by comparison with Eq. (10.4), the electric field E (assuming the potential gradient to be constant along the length of the conductor). The ratio I/A is the current per unit cross-sectional area of the conductor; it is called the *current density* and denoted J (Fig. 10.2). We can now rewrite Ohm's law as

$$E=\rho J.$$

(10.10)

This form is useful for calculating the formulae used in resistivity methods of electrical surveying. However, the quantities that are measured are V and I.

10.2.3 Types of Electrical Conduction

Electric current passes through a material by one of three different modes: by electronic, dielectric, or electrolytic conduction. Electronic (or ohmic) conduction occurs in metals and crystals, dielectric conduction in insulators, and electrolytic conduction in liquids.

Electronic conduction is typical of a metal. The free electrons in a metal have a high average speed (about 1.6×10^{6} m s^{-1} in copper). They collide with the atoms of the metal, which occupy fixed lattice sites, and bounce off in random directions. When an electric field is applied, the electrons acquire a common drift velocity, which is superposed on their random motions, so that they move at a much slower speed (about 4×10^{-5} m s^{-1} in copper) in the direction of the field. The resistivity is determined by the mean free time between collisions. If the atomic arrangement causes frequent collisions, the resistivity is high, whereas a long mean free time between collisions results in low resistivity. The energy lost in the collisions appears in the form of heat.

A form of *semi-conduction* is important in some crystals, such as the silicate minerals. The resistivity of the mineral is higher than a conductor but lower than an insulator, and it is called a *semiconductor*. Different types of semi-conduction are possible. Silicates contain fewer conduction electrons than a metal, but the electrons are not rigidly bound to atoms as in an insulator. The energy needed to liberate additional electrons from their atoms is not large, and thermal excitation is enough to allow them to take part in *electronic semi-conduction*. The liberated electron leaves a vacancy or *hole* in the valence level of the atomic structure, which behaves as a positive charge. Natural crystals also contain impurity atoms, which may have a different valency than that required by the lattice for charge balance. The impurity is a source of holes or excess electrons which take part in *impurity semi-conduction*. At high temperature, ions may detach from the lattice; they behave like ions in an electrolyte and give rise to electrical currents by *ionic semi-conduction*. A potential difference across a semiconductor produces an electric current made up of opposite flows of negative electrons and positive holes. If most of the current is carried by the negative electrons, the semiconductor is called n-type; if the positive holes predominate, the semiconductor is said to be p-type.

Dielectric conduction occurs in insulators, which contain no free electrons. Normally, the electrons are distributed symmetrically about a nucleus. However, an electric field displaces the electrons in the direction opposite to that of the field, while the heavy nucleus shifts slightly in the direction of the field. The atom or ion acquires an electric polarization and acts like an electric dipole. The net effect is to change the permittivity of the material from ε_0 to a different value ε, given by

$$\varepsilon = \varepsilon_r \varepsilon_0 . \qquad (10.11)$$

Here, ε_r is called the *relative permittivity*. When it is measured in a constant electric field, the relative permittivity is called the *dielectric constant*, κ, of the material. It is dimensionless, and has a value commonly in the range 3–80. Examples of κ for some natural materials are: air, 1.00059; mica, 3; glass, 5; sandstone, 5–12; granite, 3–19; diorite 6; basalt, 12; water 80. Dielectric effects are unimportant in constant current situations. However, in an alternating electric field the polarization changes with the frequency of the field, and thus the relative permittivity is frequency-dependent. The fluctuating polarization of the electric charge contributes to the alternating current, and so modifies the effective conductivity or resistivity. In practice, this effect depends strongly on the frequency of the inducing alternating field. The higher the frequency, the greater is the effect of dielectric conduction. Some geoelectric methods utilize signals in the audio-frequency range, where dielectric conduction is insignificant, but ground penetrating radar uses frequencies in the MHz to GHz range and depends on dielectric contrasts.

Electrolytic conduction occurs in aqueous solutions that contain free ions. The water molecule is polar (i.e., it has a permanent electric dipole moment) with a strong electric field which breaks down molecules of dissolved salts into positively and negatively charged ions. For example, in a saline solution the molecule of sodium chloride (NaCl) dissociates into separate Na^+ and Cl^- ions. The solution is called an *electrolyte*. The ions in the electrolyte are mobilized by an electric field, which causes a current to flow. Electric charge is transported by positive ions in the direction of the field and negative ions in the opposite direction. The resistivity of an electrolyte may be understood by analogy with the flow of water through a partially blocked pipe. The electric current in the electrolyte involves the physical transport of material (ions), which results in collisions with the molecules of the medium (electrolyte), causing resistance to the flow. Ionic conduction is consequently slower than electronic conduction.

10.3 Electrical Properties of the Earth

In our daily lives we experience frequent reminders of the Earth's gravity field. It is less obvious that the Earth also has an electric field. Its presence mainly becomes evident during thunderstorms, when electrical discharges take place as lightning. The Earth's electric field acts radially inward, so that the Earth behaves like a negatively charged sphere. At its surface the vertically downward electric field amounts to about 200 V m^{-1}. The atmosphere has a net positive charge, equal and opposite to that of the Earth, and resulting from the distribution of positively and negatively ionized air molecules. The charges originate from the continual bombardment of the Earth by cosmic rays.

Cosmic rays are subatomic particles with very high energy. Primary cosmic rays reach the Earth from outer space, traveling at velocities close to that of light. They consist largely of protons (hydrogen nuclei) and α-particles (helium nuclei), with lesser amounts of other ions. Their origin is still unknown. Some are emitted by the Sun at the time of solar flares, but these occur too infrequently to be the main source. This source lies elsewhere in our galaxy. It is thought that a large proportion of the galactic cosmic rays are accelerated to high speed by supernova explosions. The path of a cosmic ray is easily deflected by a magnetic field. Even the weak interstellar magnetic field is enough to disperse fast-moving cosmic rays, so that they reach the Earth equally from all directions. The incoming particles collide with nuclei in the upper atmosphere, producing showers of secondary cosmic rays, consisting of protons, neutrons, electrons, and other elementary particles. Consequently, at any given time a fraction of the molecules of the atmosphere are electrically charged. The Earth's electric field accelerates positive particles downward to the Earth's surface, where they neutralize negative surface charges. This would rapidly eliminate the negative surface charge, which is maintained by thunderstorm activity.

Thunderstorms, and the causes of lightning, are not yet fully understood. A possible scenario is the following. In a storm cloud, droplets of water vapor become electrically charged. The Earth's downward electric field may cause polarization within a droplet, with positive charge on the bottom and negative on the top. When the drop is heavy enough to fall, negative charges from molecules pushed out of its path are attracted to the bottom while fewer positive charges gather on the top. As a result, the droplet becomes negatively charged. The wind action in storm clouds causes the negative charge to accumulate at the base of the cloud, while a corresponding positive charge gathers in its upper extent. When the potential difference between the two charges exceeds the breakdown voltage of the atmosphere, a brief but powerful electric current flows. Most lightning strokes occur within the storm cloud. However, the negative charge on the base of the cloud repels the negative charge on the ground surface beneath it. Once again, if the potential difference between the cloud and the ground becomes large enough to overcome the breakdown voltage of the air, a lightning stroke ensues. This carries negative charge to the ground. In this way, the numerous daily lightning storms that occur worldwide maintain the negative charge of the Earth.

To a first approximation, the Earth may be regarded as a uniform electrical conductor. Electric charges on the surface of a conductor disperse so that the electric potential is the same at all points on the surface, i.e., it is an electrical equipotential surface. The surface potential is commonly used as the reference level for electrical potential energy and is defined to be zero. Thus a positively charged body has a positive potential difference (voltage) with respect to ground, while a negatively charged body has a negative voltage.

10.3.1 Electrical Surveying

As with other physical parameters, geoelectrical properties are utilized in both applied and general geophysics. They are exploited commercially in the search for valuable orebodies, which may be located by their anomalous electrical conductivities. Deep electrical sounding provides valuable information about the internal structure of the Earth's crust and mantle. Electrical surveys may be based on natural sources of potential and current. More commonly, they involve the detection of signals induced in subsurface conducting bodies by electric and magnetic fields generated above ground. Investigations in this category include resistivity and electromagnetic methods. These techniques have long been used in commercial geophysical surveying. In recent years they have also become important in the scientific investigation of environmental problems. The electrical techniques require the measurement of potential differences in the ground between suitably implanted electrodes. The electromagnetic techniques detect subsurface conductivity anomalies remotely; they do not need contact with the ground. As well as being employed in surface surveys they are especially suited to airborne use.

The important physical properties of rocks for electrical surveying are the permittivity (for georadar) and the resistivity (or conductivity), on which several techniques are based. Anomalies arise, for example, when a good conductor (such as a mineralized dike or orebody) is present in rocks that have higher resistivities. The resistivity contrast between orebody and host rock is often large, because the resistivities of different rocks and minerals vary widely (Fig. 10.3). In metallic ores the resistivity can be very low, but igneous rocks that contain no water can have a very high resistivity. For example, in a high-grade pyrrhotite ore ρ is of the order of 10^{-5} Ω m, while in dry marble it is around 10^{8} Ω m. The range between these extremes spans 13 orders of magnitude. Moreover, the resistivity range of any given rock type is wide and overlaps with other rock types (Fig. 10.3).

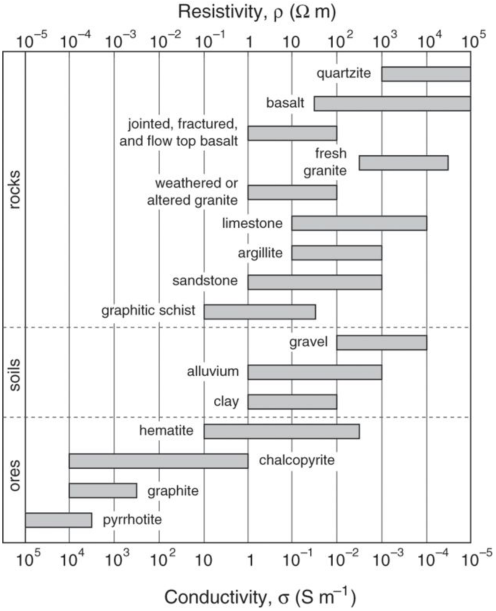


Fig. 10.3 Ranges of electrical resistivity for some common rocks, soils, and ores (data source: Ward, 1990; augmented by data from Telford *et al.*, 1990).

The resistivity of rocks is strongly influenced by the presence of groundwater, which acts as an electrolyte. This is especially important in porous sediments and sedimentary rocks. The minerals that form the matrix of a rock are generally poorer conductors than groundwater, so the conductivity of a sediment increases with the amount of groundwater it contains. This depends on the fraction of the rock that consists of pore spaces (the *porosity,* ϕ), and the fraction of this pore volume that is water-filled (the *water saturation, S*). The conductivity of the rock is proportional to the conductivity of the groundwater, which is quite variable because it depends on the concentration and type of dissolved minerals and salts it contains. These observations are summarized in an empirical formula, called Archie's law, for the resistivity ρ of the rock:

$$\rho = \frac{a}{\phi^m S^n} \rho_w. \tag{10.12}$$

By definition, ϕ and S are fractions between 0 and 1, ρ_w is the resistivity of the groundwater, and the parameters a, m, and n are empirical constants that have to be determined for each case. Generally, $0.5 \leq a \leq 2.5$, $1.3 \leq m \leq 2.5$ and $n \approx 2$.

10.4 Natural Potentials and Currents

Electrical investigations of natural electrical properties are based on the measurement of the voltage between a pair of electrodes implanted in the ground. Natural differences in potential occur in relation to subsurface bodies that create their own electric fields. The bodies act like simple voltaic cells; their potential arises from electrochemical action. Natural currents (called *telluric currents*) flow in the crust and mantle of the Earth. They are induced electromagnetically by electric currents in the *ionosphere* (described in Section 11.2.3.2). In studying natural potentials and currents, the scientist has no control over the source of the signal. This restricts the interpretation, which is mostly only qualitative. The natural methods are not as useful as controlled induction methods, such as resistivity and electromagnetic techniques, but they are inexpensive and fast.

10.4.1 Self-Potential (Spontaneous Potential)

A potential that originates spontaneously in the ground is called a *self-potential* (or *spontaneous* potential). Some self-potentials are due to man-made disturbances of the environment, such as buried electrical cables, drainage pipes, or waste disposal sites. They are important in the study of environmental problems. Other self-potentials are natural effects due to mechanical or electrochemical action. In every case the groundwater plays a key role by acting as an electrolyte.

Some self-potentials have a mechanical origin. When an electrolyte is forced to flow through a narrow pipe, a potential difference (voltage) may arise between the ends of the pipe. Its amplitude depends on the electrical resistivity and viscosity of the electrolyte, and on the pressure difference that causes the flow. The voltage is due to differences in the *electrokinetic* or *streaming potential,* which in turn is influenced by the interaction between the liquid and the surface of the solid (an effect called the *zeta-potential*). The voltage can be positive or negative and may amount to some hundreds of millivolts. This type of self-potential can be observed in conjunction with seepage of water from dams, or the flow of groundwater through different lithological units.

Most self-potentials have an electrochemical origin. For example, if the ionic concentration in an electrolyte varies with location, the ions tend to diffuse through the electrolyte so as to equalize the concentration. The diffusion is driven by an electric *diffusion potential,* which depends on the temperature as well as the difference in ionic concentration. When a metallic electrode is inserted in the ground, the metal reacts electrochemically with the electrolyte (i.e., groundwater), causing a contact potential. If two identical electrodes are inserted in the ground, variations in concentration of the electrolyte cause different electrochemical reactions at each electrode. A potential difference arises, called the *Nernst potential.* The combined diffusion and Nernst potentials are called the *electrochemical* self-potential. It is temperature sensitive and may be either positive or negative, amounting to at most a few tens of millivolts.

The self-potentials that originate by the above mechanisms are attracting increased attention in environmental and engineering situations. However, in the exploration for subsurface regions of mineralization they are often smaller than the potentials associated with orebodies and are classified accordingly as “background potentials.” The self-potential associated with an orebody is called its “mineralization potential.” Self-potential anomalies across orebodies are invariably negative, amounting usually to a few hundred millivolts. They are most commonly associated with sulfide ores, such as pyrite, pyrrhotite, and chalcopyrite, but also with graphite and some metallic oxides.

The origin of the mineralization type of self-potential is still obscure, despite decades of applied investigations. At one time it was thought that the effect arose from galvanic action. This occurs when dissimilar metal electrodes are placed in an electrolyte. Unequal contact potentials are formed between the metals and the electrolyte, giving rise to a potential difference between the electrodes. According to this model an orebody behaves like a simple voltaic cell, with groundwater acting as the electrolyte. It was believed that oxidation of the part of the orebody above the water table produced a potential difference between the upper and lower parts, causing a spontaneous electric polarization of the body. Oxidation involves the addition of electrons, so the top of the orebody becomes negatively charged, explaining the observed negative anomalies. Unfortunately, this simple model does not explain many of the observed features of self-potential anomalies and has proved to be untenable.

Another mechanism for self-potential depends on variations in oxidation (redox) potential with depth (Fig. 10.4). The ground above the water table is more accessible to oxygen than the submersed part, so moisture above the water table contains more oxidized ions than that below it. An electrochemical reaction takes place at the surface between the orebody and the host rock above the water table. It results in reduction of the oxidized ions in the adjacent solution. An excess of negative ions appears above the water table. A simultaneous reaction between the submersed part of the orebody and the groundwater causes oxidation of the reduced ions present in the groundwater. This produces excess positive ions in the solution and liberates electrons at the surface of the orebody, which acts as a conductor connecting the two half-cells. Electrons flow from the deep part to the shallow part of the orebody. Outside the orebody, positive ions move from bottom to top along the electric field lines. The equipotential surfaces are normal to the field lines. The self-potential is measured where they intersect the ground surface (Fig. 10.4).

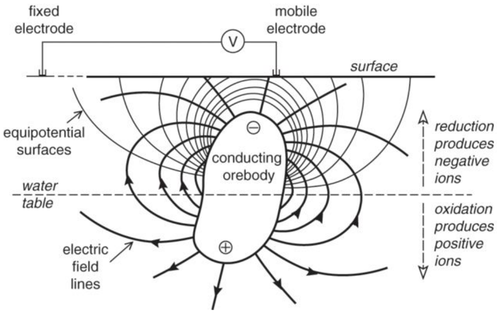

fixed electrode
mobile electrode
V
surface
equipotential surfaces
⊖
conducting orebody
reduction produces negative ions
water table
oxidation produces positive ions
electric field lines
⊕

Fig. 10.4 A schematic model of the origin of the self-potential anomaly of an orebody. The mechanism depends on differences in oxidation potential above and below the water table.

The redox model is inadequate for the same reason as the galvanic model; it fails to account for many of the observed features of self-potential anomalies. In particular, the association of self-potential models with the water table has been cast into doubt. Moreover, sulfide orebodies appear to persist for geological lengths of time, so that a mechanism involving permanent flow of charge appears unlikely. Self-potential is a feature of a stable system that is perturbed by making an electrical connection between the host rock and the sulfide conductor through the inserted electrodes and their connecting wire. The observed potential difference appears to be due to the difference in oxidation potential between the locations of the measurement electrodes, one inside and the other outside the zone of mineralization.

10.4.2 Self-Potential Surveying

The equipment needed for a self-potential survey is very simple. It consists of a sensitive high-impedance digital voltmeter to measure the natural potential difference between two electrodes implanted in the ground. Simple metal stakes are inadequate as electrodes. Electrochemical reactions take place between the metal and moisture in the ground, causing the build-up of spurious charges on the electrodes, which can falsify or obscure the small natural self-potentials. To avoid or minimize this effect, *non-polarizable electrodes* are used. Each electrode consists of a metal rod submersed in a saturated solution of its own salt; a common arrangement is a copper rod in copper sulfate solution. The combination is contained in a ceramic pot that allows the electrolyte to leak slowly through its porous walls, thereby making electrical contact with the ground.

Two field methods are in common use (Fig. 10.5). The gradient method employs a fixed separation between the electrodes, of the order of 10 m apart. The potential difference is measured between the electrodes, then the pair is moved forward along the survey line until the trailing electrode occupies the location previously occupied by the leading electrode. The total potential at a measurement station relative to a starting point outside the study area is found by summing the incremental potential differences. Some electrode polarization is unavoidable, even with non-polarizable electrodes. This gives rise to a small error in each measurement; these add up to a cumulative error in the total potential. The polarization effects can sometimes be reduced by interchanging the leading and trailing electrodes. In this “leapfrog” technique, the leading electrode for one measurement is kept in place and becomes the trailing electrode for the next measurement; meanwhile the previous trailing electrode is moved ahead to become the leading electrode. Cumulative error is the most serious disadvantage of the fixed electrode configuration. A practical advantage of the technique is that only a short length of connecting wire must be moved along with the electrodes.

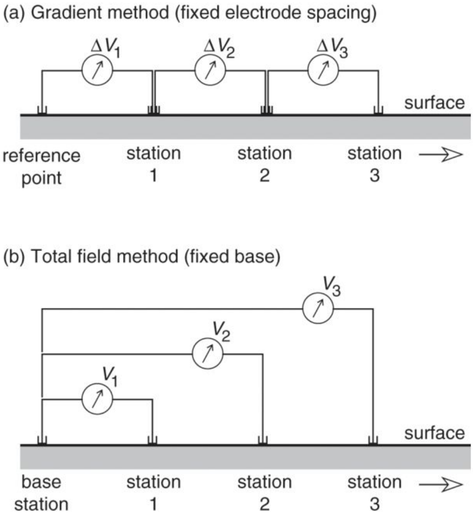

(a) Gradient method (fixed electrode spacing)
ΔV1
ΔV2
ΔV3
surface
reference point
station 1
station 2
station 3
(b) Total field method (fixed base)
V3
V2
V1
surface
base station
station 1
station 2
station 3

Fig. 10.5 The field techniques of measuring self-potential by (a) the gradient method and (b) the total field method. The total potential *V* at a station in the gradient method is found by summing the previous potential differences Δ*V*; in the total field method *V* is measured directly.

The total field method utilizes a fixed electrode at a base station outside the area of exploration and a mobile measuring electrode. With this method the total potential is measured directly at each station. The wire connecting the electrodes has to be long enough to allow good coverage of the area of interest. This necessitates a long wire that must be wound or unwound on a reel for each measurement station. However, the total field method results in smaller cumulative error than the gradient method. It allows more flexibility in placing the mobile electrode and usually gives data of better quality. Hence, the total field method is usually preferred except in difficult terrain.

The surveying procedure with each technique consists of measuring potential at discrete stations along a profile. As in gravity and magnetic surveys, the data are mapped (Fig. 10.6) and interpretations of anomalies are based on their geometry. Methods used to interpret self-potential anomalies are often qualitative or are based on simple geometric models. Visual inspection of mapped anomalies may reveal trends related to elongation of the orebody; crowding of contour lines can indicate its orientation. Profiles plotted in known directions across the anomaly can be compared with curves generated from simple models of the source. For example, a polarized sphere may be used to model the source of approximately circular anomalies, while a horizontal line source (or polarized cylinder) may be used to model an elongate anomaly. A common and effective method is to model self-potential anomalies with point sources; complex anomalies are modeled with combinations of sources and sinks.

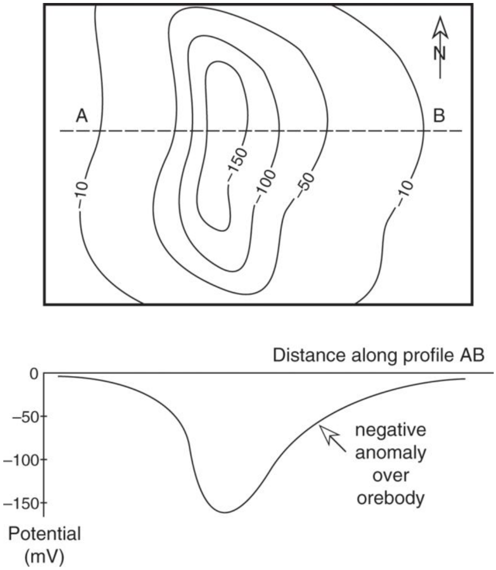


Fig. 10.6 Hypothetical contour lines of a negative self-potential anomaly over an orebody; the asymmetry of the anomaly along the profile AB suggests that the orebody dips toward A.

10.4.3 Telluric Currents

Ultraviolet radiation from the Sun ionizes molecules of air in the thin upper atmosphere of the Earth. The ions accumulate in several layers, forming the ionosphere (see Section 11.2.3.2) at altitudes between about 80 km and 1500 km above the Earth's surface. Electric currents in the ionosphere arise from systematic motions of the ions, which are affected by various factors such as the daily and monthly tides, seasonal variations in insolation, and the periodic fluctuation in ionization related to the 11-year sunspot cycle. The currents produce varying magnetic fields with the same frequencies, which are observed at the surface of the Earth and can be analyzed from long-term continuous records of the geomagnetic field. The ionospheric effects show up in the energy spectrum of the geomagnetic field as distinct peaks representing periods that range from fractions of a second (geomagnetic *pulsations*) to several years (Fig. 10.7). The magnetic fields induce fluctuating electric currents, called *telluric currents*, that flow in horizontal layers in the crust and mantle. The current pattern consists of several huge whorls, thousands of kilometers across, which remain fixed with respect to the Sun and thus move around the Earth as it rotates.

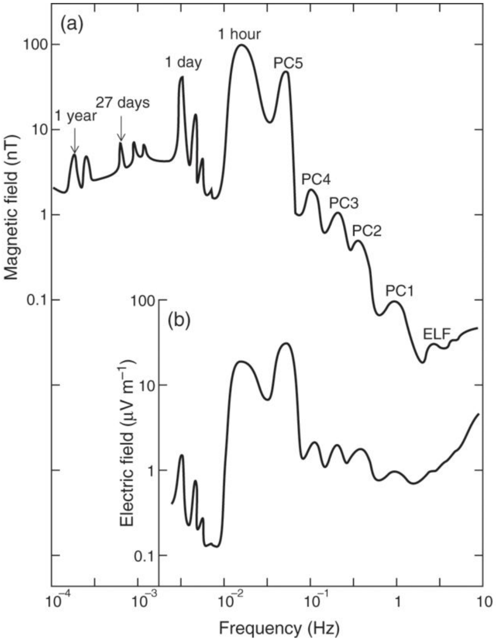


Fig. 10.7 (a) The frequency spectrum of natural variations in the horizontal intensity of the geomagnetic field, and (b) the corresponding spectrum of induced electric field fluctuations, computed for a model Earth with uniform resistivity 20 Ω m.

Source: P.H. Serson, Instrumentation for induction studies on land, *Phys. Earth Planet. Int.* 7, 313–322, 1973.

The distribution of telluric current density depends on the variation of resistivity in the horizontal conducting layers. At shallow crustal depths the lines of current flow are disturbed by subsurface structures which cause contrasts in resistivity. These could arise from geological structures or the presence of mineralized zones. Consider, for example, a buried anticline which has a highly resistive rock (such as granite) as its core and is overlain by a conducting layer of porous sedimentary rocks saturated with groundwater. The horizontal flow of telluric current across the anticline chooses the less-resistive path through the conducting sediments. The current lines bunch together over the axis of the anticline, increasing the horizontal current density (Fig. 10.8). The equipotential surfaces normal to the current lines intersect the ground surface, where potential differences can be measured with a high-impedance voltmeter.

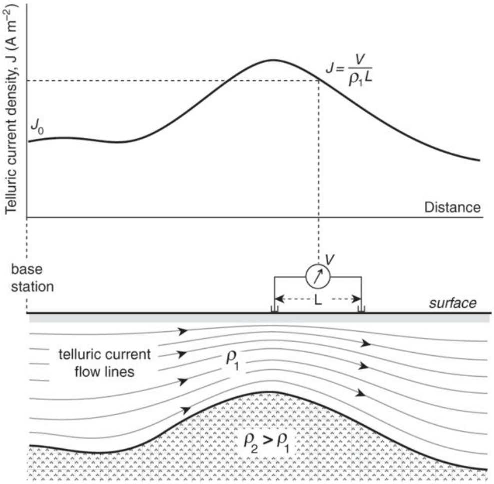

Telluric current density, J (A m⁻²)
J₀
J = V/ρ₁L
Distance
base station
V
L
surface
telluric current flow lines
ρ₁
ρ₂ > ρ₁

Fig. 10.8 Telluric current lines are deflected by changes in thickness of a conducting layer over a more resistive structure (*bottom*). The telluric current density (*top*) is obtained from the voltage measured between a pair of fixed-separation electrodes at the surface.

Source: E.S. Robinson and C. Çoruh, *Basic Exploration Geophysics*, 562 pp., New York: John Wiley, 1988.

The field equipment for measuring telluric current density is simple. The sensors are a pair of non-polarizable electrodes with a fixed separation *L* of the order of 10–100 m. The potential difference *V* between the electrodes is measured with a high-impedance voltmeter. The electric field *E* at a point mid-way between the electrodes can be assumed to be *V*/*L*. Using Ohm's law (Eq. 10.10) and assuming that the telluric current flows in conducting rock layer with resistivity ρ_1, the telluric current density *J* at each measurement station along a profile is given by

$$J = \frac{V}{\rho_1 L}. \tag{10.13}$$

The direction of the telluric current is not known, so two pairs of electrodes oriented perpendicular to each other are used. One pair is aligned north–south and the other east–west. Telluric currents vary unpredictably with time, but they change only slowly within a homogeneous region. To keep track of the temporal changes, an orthogonal pair of electrodes is set up at a fixed base station outside the area to be explored. Another orthogonal pair is moved across the survey area. The potential differences across each electrode pair in the mobile and base arrays are recorded simultaneously for several minutes at each measurement station. Correlation of the records allows removal of the temporal changes in direction and intensity of the telluric currents.

The deflection of telluric current by a resistive subsurface structure as shown in Fig. 10.8 is greatly idealized. It assumes an infinite resistivity ρ_2 in the core of the anticline. In practice, the current is not completely diverted through the better-conducting layer; part flows through the more resistive layer as well. Thus we cannot assume that the resistivity ρ_1 in Eq. (10.13) corresponds to the good conductor. Rather, it represents some undefined mixture of the values ρ_1 and ρ_2. It is not the true resistivity of either layer, but the *apparent resistivity* of the measurement.

10.5 Resistivity Surveying

The large contrast in resistivity between orebodies and their host rocks (see Fig. 10.3) is exploited in electrical resistivity prospecting, especially for minerals that occur as good conductors. Representative examples are the sulfide ores of iron, copper, and nickel. Electrical resistivity surveying is also an important geophysical technique in environmental applications. For example, due to the good electrical conductivity of groundwater the resistivity of a sedimentary rock is much lower when it is waterlogged than in the dry state.

Instead of relying on natural currents, two electrodes are used to supply a controlled electrical current to the ground. As in the telluric method, the lines of current flow adapt to the subsurface resistivity pattern so that the potential difference between equipotential surfaces can be measured where they intersect the ground surface, using a second pair of electrodes. A simple direct current can cause charges to accumulate on the potential electrodes, which results in spurious signals. A common practice is to commutate the direct current so that its direction is reversed every few seconds; alternatively a low-frequency alternating current may be used. In multi-electrode investigations the current electrode-pair and potential electrode-pair are usually interchangeable.

10.5.1 Potential of a Single Electrode

Consider the flow of current around an electrode that introduces a current *I* at the surface of a uniform half-space (Fig. 10.9a). The point of contact acts as a current source, from which the current disperses outward. The electric field lines are parallel to the current flow and normal to the equipotential surfaces, which are hemispherical in shape. The current density *J* is equal to *I* divided by the surface area, which is $2\pi r^2$ for a hemisphere of radius *r*. The electric field *E* at distance *r* from the input electrode is obtained from Ohm's law (Eq. 10.10):

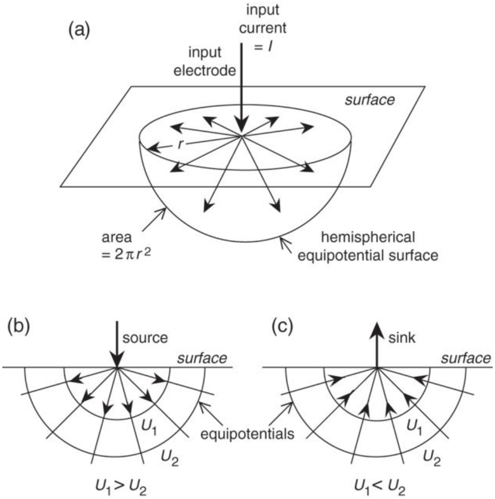

(a)
input current = I
input electrode
surface
r
area = 2πr²
hemispherical equipotential surface
(b)
source
surface
U₁
U₂
U₁ > U₂
equipotentials
(c)
sink
surface
U₁
U₂
U₁ < U₂

Fig. 10.9 Electric field lines and equipotential surfaces around a single electrode at the surface of a uniform half-space: (a) hemispherical equipotential surfaces, (b) radially outward field lines around a source, and (c) radially inward field lines around a sink.

$$E = \rho J = \rho \frac{I}{2\pi r^2}. \tag{10.14}$$

Putting this expression in Eq. (10.4) yields the electric potential *U* at distance *r* from the input electrode:

$$\frac{dU}{dr} = -\rho\frac{I}{2\pi r^2}$$

$$U = \rho\frac{I}{2\pi r}. \tag{10.15}$$

If the ground is a uniform half-space, the electric field lines around a *source* electrode, which supplies current to the ground, are radially outward (Fig. 10.9b). Around a *sink* electrode, where current flows out of the ground, the field lines are radially inward (Fig. 10.9c). The equipotential surfaces around a source or sink electrode are hemispheres, if we regard the electrode in isolation. The potential around a source is positive and diminishes as $1/r$ with increasing distance. The sign of I is negative at a sink, where the current flows out of the ground. Thus, around a sink the potential is negative and increases (becomes less negative) as $1/r$ with increasing distance from the sink. We can use these observations to calculate the potential difference between a second pair of electrodes at known distances from the source and sink.

10.5.2 The General Four-Electrode Method

Consider an arrangement consisting of a pair of current electrodes and a pair of potential electrodes (Fig. 10.10). The current electrodes A and B act as source and sink, respectively. At the detection electrode C the potential due to the source A is + $\rho I/(2\pi\, r_{AC})$, while the potential due to the sink B is – $\rho I/(2\pi\, r_{CB})$. The combined potential at C is

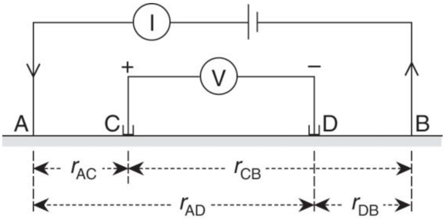


Fig. 10.10 General four-electrode configuration for resistivity measurement, consisting of a pair of current electrodes (A, B) and a pair of potential electrodes (C, D).

$$U_C = \frac{\rho I}{2\pi}\left(\frac{1}{r_{AC}} - \frac{1}{r_{CB}}\right). \tag{10.16}$$

Similarly, the resultant potential at D is

$$U_D = \frac{\rho I}{2\pi}\left(\frac{1}{r_{AD}} - \frac{1}{r_{DB}}\right). \tag{10.17}$$

The potential difference measured by a voltmeter connected between C and D is

$$V = \frac{\rho I}{2\pi}\left\{\left(\frac{1}{r_{AC}} - \frac{1}{r_{CB}}\right) - \left(\frac{1}{r_{AD}} - \frac{1}{r_{DB}}\right)\right\}. \tag{10.18}$$

All quantities in this equation can be measured at the ground surface except the resistivity, which is given by

$$\rho = 2\pi \frac{V}{I} \left\{ \left(\frac{1}{r_{\mathrm{AC}}} - \frac{1}{r_{\mathrm{CB}}} \right) - \left(\frac{1}{r_{\mathrm{AD}}} - \frac{1}{r_{\mathrm{DB}}} \right) \right\}^{-1}. \tag{10.19}$$

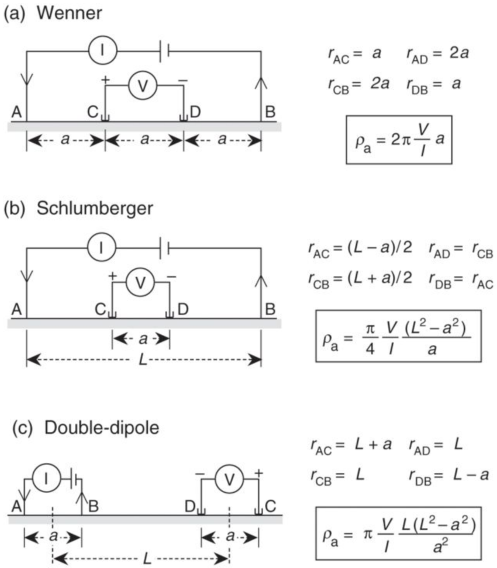

(a) Wenner
I
V
+
−
A
C
D
B
a
a
a
$r_{AC} = a$
$r_{AD} = 2a$
$r_{CB} = 2a$
$r_{DB} = a$
$\rho_a = 2\pi \frac{V}{I} a$
(b) Schlumberger
I
V
+
−
A
C
D
B
a
L
$r_{AC} = (L-a)/2$
$r_{AD} = r_{CB}$
$r_{CB} = (L+a)/2$
$r_{DB} = r_{AC}$
$\rho_a = \frac{\pi}{4} \frac{V}{I} \frac{(L^2-a^2)}{a}$
(c) Double-dipole
I
V
−
+
A
B
D
C
a
a
L
$r_{AC} = L + a$
$r_{AD} = L$
$r_{CB} = L$
$r_{DB} = L - a$
$\rho_a = \pi \frac{V}{I} \frac{L(L^2-a^2)}{a^2}$

Fig. 10.11 Special geometries of current and potential electrodes for (a) Wenner, (b) Schlumberger, and (c) double-dipole configurations.

10.5.3 Special Electrode Configurations

The general formula for the resistivity measured by a four-electrode method is simpler for some special geometries of the current and potential electrodes. The most commonly used configurations are the Wenner, Schlumberger, and double-dipole arrangements. In each configuration the four electrodes are collinear, but their geometries and spacings are different.

In the *Wenner configuration* (Fig. 10.11a), the current and potential electrode pairs have a common mid-point and the distances between adjacent electrodes are equal, so that $r_{AC} = r_{DB} = a$, and $r_{CB} = r_{AD} = 2a$. Inserting these values in Eq. (10.19) gives

$$\rho = 2\pi \frac{V}{I} \left\{ \left(\frac{1}{a} - \frac{1}{2a} \right) - \left(\frac{1}{2a} - \frac{1}{a} \right) \right\}^{-1},$$

(10.20)

$$\rho = 2\pi a \frac{V}{I}.$$

(10.21)

In the *Schlumberger configuration* (Fig. 10.11b) the current and potential pairs of electrodes often also have a common mid-point, but the distances between adjacent electrodes differ. Let the separations of the current and potential electrodes be L and a, respectively. Then $r_{AC} = r_{DB} = (L - a)/2$ and $r_{AD} = r_{CB} = (L + a)/2$. Substituting in the general formula, we get

$$\rho = 2\pi \frac{V}{I} \left\{ \left(\frac{2}{L-a} - \frac{2}{L+a} \right) - \left(\frac{2}{L+a} - \frac{2}{L-a} \right) \right\}^{-1}$$
$$= \frac{\pi}{4} \frac{V}{I} \left(\frac{L^2 - a^2}{a} \right).$$

(10.22)

In this configuration the separation of the current electrodes is kept much larger than that of the potential electrodes (L ≫ a). Under these conditions, Eq. (10.22) simplifies to

$$\rho = \frac{\pi}{4}\frac{V}{I}\left(\frac{L^2}{a}\right). \tag{10.23}$$

In the *double-dipole configuration* (Fig. 10.11c) the spacing of the electrodes in each pair is *a*, while the distance between their mid-points is *L*, which is generally much larger than *a*. Note that detection electrode D is defined as the potential electrode closer to current sink B. In this case $r_{AD} = r_{BC} = L$, $r_{AC} = L + a$, and $r_{DB} = L - a$. The measured resistivity is

$$\rho = 2\pi\frac{V}{I}\left\{\left(\frac{1}{L+a} - \frac{1}{L}\right) - \left(\frac{1}{L} - \frac{1}{L-a}\right)\right\}^{-1}, \tag{10.24}$$

$$\rho = \pi\frac{V}{I}\left(\frac{L(L^2 - a^2)}{a^2}\right). \tag{10.25}$$

Two modes of investigation can be used with each electrode configuration. The Wenner configuration is best adapted to *lateral profiling*. The assemblage of four electrodes is displaced stepwise along a profile while maintaining constant values of the inter-electrode distances corresponding to the configuration employed. The separation of the current electrodes is chosen so that the current flow is maximized in depths where lateral resistivity contrasts are expected. Results from a number of profiles may be compiled in a resistivity map of the region of interest. The regional survey reveals the horizontal variations in resistivity within an area at a particular depth.

It is best suited to locating steeply dipping contacts between rocks with a strong resistivity contrast and good conductors such as mineralized dikes, which may be potential orebodies.

In *vertical electrical sounding* (VES), the goal is to observe the variation of resistivity with depth. The technique is best adapted to determining depth and resistivity for flat-lying layered rock structures, such as sedimentary beds, or the depth to the water table. The Schlumberger configuration is most commonly used for VES investigations. The mid-point of the array is kept fixed while the distance between the current electrodes is progressively increased. This causes the current lines to penetrate to ever greater depths, depending on the vertical distribution of conductivity.

10.5.4 Current Distribution

The current pattern in a uniform half-space extends laterally on either side of the profile line. Viewed from above, the current lines bulge outward between source and sink with a geometry similar to that shown in Fig. 10.1c. In a vertical section the current lines resemble half of a dipole geometry. In three dimensions the current can be visualized as flowing through tubes that fatten as they leave the source and narrow as they converge toward the sink. Fig. 10.12 shows the flow pattern of the current in a vertical section through the "tubes" in a uniform half-space.

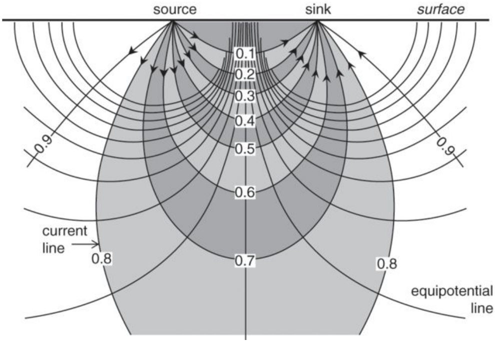


Fig. 10.12 Cross-section of current "tubes" and equipotential surfaces between a source and sink; numbers on the current lines indicate the fraction of current flowing above the line.

Source: E.S. Robinson and C. Çoruh, *Basic Exploration Geophysics*, 562 pp., New York: John Wiley, 1988.

In order to evaluate the depth penetration of current in a uniform half-space, we define orthogonal Cartesian coordinates with the x-axis parallel to the profile and the z-axis vertical (Fig. 10.13a). Let the spacing of the current electrodes be L and the resistivity of the half-space be ρ. The horizontal electric field E_x at (x, y, z) is

$$E_x = -\frac{\partial U}{\partial x} = -\frac{\partial}{\partial x}\left\{\frac{\rho I}{2\pi}\left(\frac{1}{r_1} - \frac{1}{r_2}\right)\right\},$$

(10.26)

where $r_1 = (x^2 + y^2 + z^2)^{1/2}$ and $r_2 = ((L-x)^2 + y^2 + z^2)^{1/2}$. Differentiating and using Ohm's law (Eq. 10.10) gives the horizontal current density J_x at (x, y, z):

$$J_x = \frac{I}{2\pi}\left(\frac{x}{r_1^3} + \frac{L-x}{r_2^3}\right).$$

(10.27)

If (x, y, z) is on the vertical plane mid-way between the current electrodes, $x = L/2$, $r_1 = r_2$ and the current density is given by

$$J_x = \frac{IL}{2\pi}\frac{1}{((L/2)^2 + y^2 + z^2)^{3/2}}.$$

(10.28)

The horizontal current $\mathrm{d}I_x$ across an element of area (dy dz) in the median vertical plane is $\mathrm{d}I_x = J_x\,\mathrm{d}y\,\mathrm{d}z$. The fraction of the input current I that flows across the median plane above a depth z is obtained by integration:

$$\frac{I_x}{I} = \frac{L}{2\pi}\int_0^z \mathrm{d}z \int_{-\infty}^{+\infty} \frac{\mathrm{d}y}{((L/2)^2 + y^2 + z^2)^{3/2}},$$

(10.29)

$$\frac{I_x}{I} = \frac{L}{\pi}\int_0^z \frac{\mathrm{d}z}{((L/2)^2 + z^2)},$$

(10.30)

$$\frac{I_x}{I} = \frac{2}{\pi}\tan^{-1}\left(\frac{2z}{L}\right). \tag{10.31}$$

Equation (10.31) shows that I_x depends upon the current-electrode spacing L (Fig. 10.13b). Half the current crosses the plane above a depth $z = L/2$, and almost 90% passes above the depth $z = 3L$. The fraction of current between any two depths is found from the difference in the fractions above each depth calculated with Eq. (10.31).

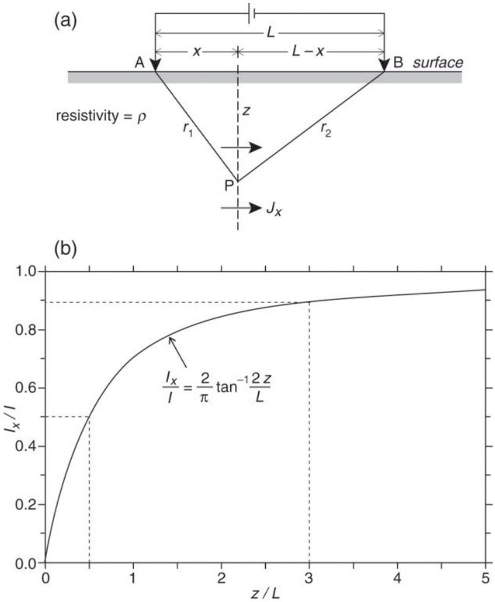


Fig. 10.13 (a) Geometry for determining current density in uniform ground below two electrodes, and (b) fraction of current (I_x/I) that flows above depth z across the median plane between current electrodes with spacing.

Source: W.M. Telford, L.P. Geldart and R.E. Sheriff, *Applied Geophysics*, 770 pp., Cambridge: Cambridge University Press, 1990.

10.5.5 Apparent Resistivity

In the idealized case of a perfectly uniform conducting half-space, the current flow lines resemble a dipole pattern (Fig. 10.12), and the resistivity determined with a four-electrode configuration is the true resistivity of the half-space. But in real situations the resistivity is determined by different lithologies and geological structures and so may be very inhomogeneous. This complexity is not taken into account when measuring resistivity with a four-electrode method, which assumes that the ground is uniform. The result of such a measurement is the *apparent resistivity* of an equivalent uniform half-space and generally does not represent the true resistivity of any part of the ground.

Consider a horizontally layered structure in which a layer of thickness d and resistivity ρ_1 overlies a conducting half-space with a lower resistivity ρ_2 (Fig. 10.14). If the current electrodes are close together, so that $L \ll d$, all or most of the current flows in the more resistive upper layer, so that the measured resistivity is close to the true value of the upper layer, ρ_1. With increasing separation of the current electrodes the depth reached by the current lines increases. Proportionally more current flows in the less resistive layer, so the measured resistivity decreases. Conversely, if the upper layer is a better conductor than the lower layer, the apparent resistivity increases with increasing electrode spacing. When the electrode separation is much larger than the thickness of the upper layer ($L \gg d$) the measured resistivity is close to the value ρ_2 of the bottom layer. Between the extreme situations the apparent resistivity determined from the measured current and voltage is not related simply to the true resistivity of either layer.

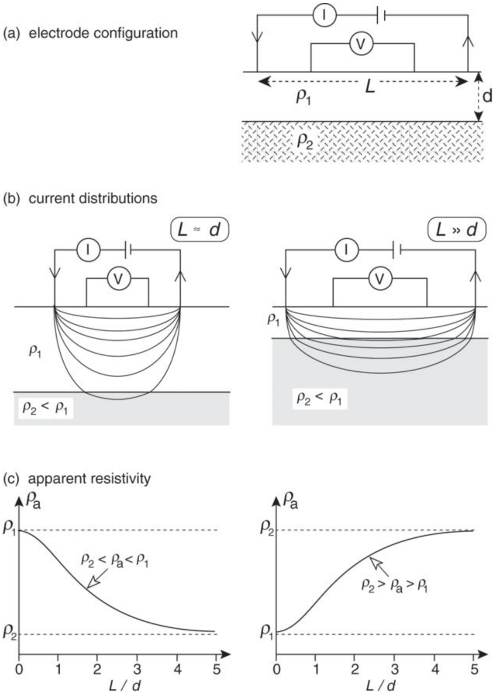

(a) electrode configuration
I
V
L
d
ρ1
ρ2
(b) current distributions
L ≈ d
L » d
I
V
ρ1
ρ2 < ρ1
(c) apparent resistivity
ρa
ρ1
ρ2
ρ2 < ρa < ρ1
ρ2 > ρa > ρ1
0
1
2
3
4
5
L / d

Fig. 10.14 (a) Parameters of the four-electrode arrangement; (b) distribution of current lines in a two-layer ground with resistivities ρ_1 and ρ_2 ($\rho_1 > \rho_2$); and (c) the variation of apparent resistivity as the current electrode spacing is varied for the two cases of $\rho_1 > \rho_2$ and $\rho_1 < \rho_2$.

10.5.6 Vertical Electrical Sounding

A two-layer situation is encountered often in electrical prospecting – for example, when a conducting overburden overlies a resistive basement. It is also common in environmental applications, when the conducting water table lies under drier, more resistive soil or rocks. Before the advent of portable computers, two-layer cases were interpreted with the aid of *characteristic curves*. These theoretical curves, calculated for a particular four-electrode array, take into account the change of depth penetration when current lines cross the boundary to a layer with different resistivity. The electrical boundary conditions require continuity of the component of current density *J* normal to the interface and of the component of electric field *E* tangential to the interface. At a boundary the current lines behave like optical or seismic rays, and are guided by similar laws of reflection and refraction. For example, if θ is the angle between a current line and the normal to the interface, the electrical "law of refraction" is

$$\frac{\tan\theta_1}{\tan\theta_2}=\frac{\rho_2}{\rho_1}. \tag{10.32}$$

In a set of characteristic curves the apparent resistivity ρ_a is normalized by the resistivity ρ_1 of the upper layer and the electrode spacing is expressed as a multiple of the layer thickness. The shape of the curve of apparent resistivity versus electrode spacing depends on the resistivity contrast between the two layers, and a family of characteristic curves is calculated for different ratios of ρ_2/ρ_1 (Fig. 10.15). The resistivity contrast is conveniently expressed by a *k*-factor defined as

$$k=\frac{\rho_2-\rho_1}{\rho_2+\rho_1}. \tag{10.33}$$

The *k*-factor ranges between −1 and +1 as the resistivity ratio ρ_2/ρ_1 varies between 0 and ∞ . The characteristic curves, drawn as full logarithmic plots on a transparent overlay, are compared graphically with the field data to find the best-fitting characteristic curve. The comparison yields the resistivities ρ_1 and ρ_2 of the upper and lower layers, respectively, and the layer thickness, *d*.

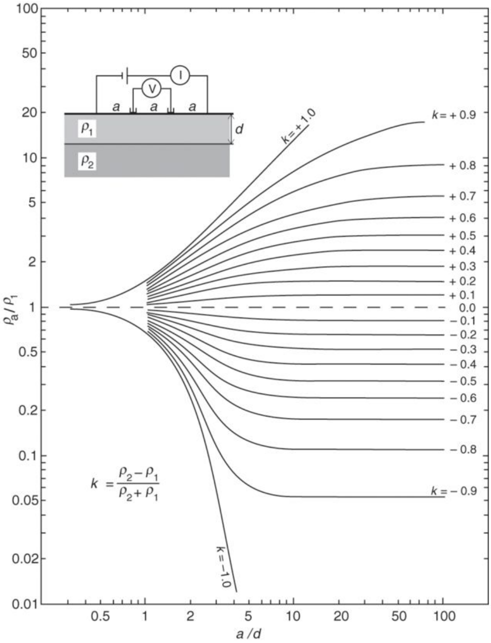


Fig. 10.15 Characteristic curves of apparent resistivity for a two-layer structure using the Wenner array; parameters are defined in the inset.

Although characteristic curves can also be computed for the interpretation of structures with multiple horizontal layers, modern VES analyses take advantage of the flexibility offered by small computers with graphic outputs on which the apparent resistivity curves can be visually assessed. The first step in the analysis consists of classifying the shape of the vertical sounding profile.

The apparent resistivity curve for a three-layer structure generally has one of four typical shapes, determined by the vertical sequence of resistivities in the layers (Fig. 10.16). The type K curve rises to a maximum then decreases, indicating that the intermediate layer has higher resistivity than the top and bottom layers. The type H curve shows the opposite effect; it falls to a minimum then increases again due to an intermediate layer that is a better conductor than the top and bottom layers. The type A curve may show some changes of gradient, but the apparent resistivity generally increases continuously with increasing electrode separation, indicating that the true resistivities increase with depth from layer to layer. The type Q curve exhibits the opposite effect; it decreases continuously along with a progressive decrease of resistivity with depth.

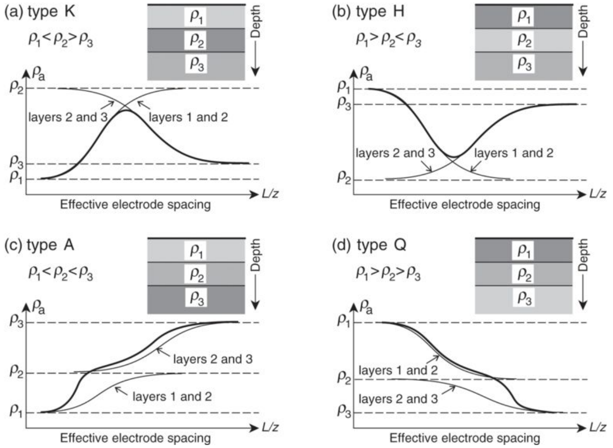


Fig. 10.16 The four common shapes of apparent resistivity curves for a layered structure consisting of three horizontal layers.

Once the observed resistivity profile has been identified as of K, H, A, or Q type, the next step is equivalent to one-dimensional *inversion* of the field data. The technique involves iterative procedures that would be very time-consuming without a fast computer. The method assumes the equations for the theoretical response of a multi-layered ground. Each layer is characterized by its thickness and resistivity, each of which must be determined. A first estimate of these parameters is made for each layer and the predicted curve of apparent resistivity versus electrode spacing is computed. The discrepancies between the observed and theoretical curves are then determined point by point. The layer parameters used in the governing equations are next adjusted, and the calculation is repeated with the corrected values, giving a new predicted curve to compare with the field data. Using modern computers the procedure can be reiterated rapidly until the discrepancies are smaller than a pre-determined value.

The inversion method is equivalent to matching automatically the observed and theoretical curves. A one-dimensional analysis accommodates only the variations of resistivity and layer thickness with depth. The response of a vertically layered structure has an analytical solution, so efficient inversion algorithms can be established. In recent years, procedures have been proposed that also take into account lateral heterogeneities. The response of two- or three-dimensional structures must be approximated by a numerical solution, based on the finite-difference or finite-element techniques. The number of unknown quantities increases, as do the computational difficulties of the inversion.

10.5.7 Induced Polarization

If commutated direct current is used in a four-electrode resistivity survey, the sequence of positive and negative flow may be interspersed with periods when the current is off. The inducing current then has a box-like appearance (Fig. 10.17a). When the current is interrupted, the voltage across the potential electrodes does not drop immediately to zero. After an initial abrupt drop to a fraction of its steady-state value it decays slowly for several seconds (Fig. 10.17b). Conversely, when the current is switched on, the potential rises suddenly at first and then gradually approaches the steady-state value. The slow decay and growth of part of the signal are due to *induced polarization*, which results from two similar effects related to the rock structure: membrane polarization and electrode polarization.

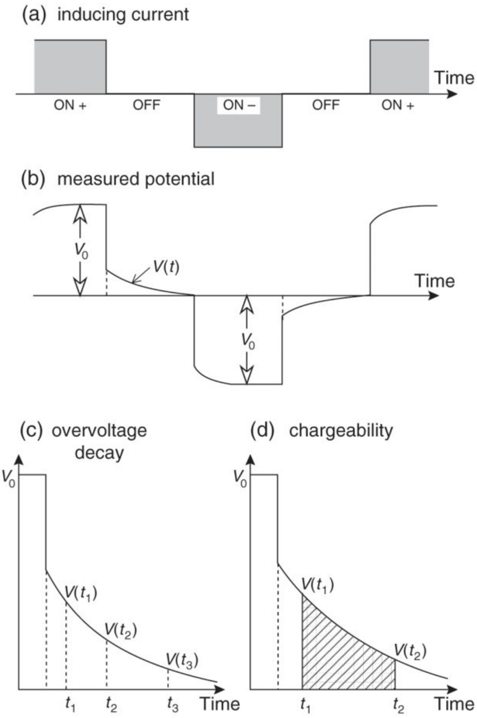


Fig. 10.17 (a) Illustration of the IP-related decay of potential after interruption of the primary current. (b) Effect of the IP decay time on the potential waveform for a square-wave input current.

Membrane polarization is a feature of electrolytic conduction. It arises from differences in the ability of ions in pore fluids to migrate through a porous rock. The minerals in a rock generally have a negative surface charge and thus attract positive ions in the pore fluid. They accumulate on the grain surface and extend into the adjacent pores, partially blocking them. When an external voltage is applied, positive ions can pass through the "cloud" of positive charge but negative ions accumulate, unless the pore size is large enough to allow them to bypass the blockage. The effect is like a membrane, which selectively allows the passage of one type of ion. It causes temporary accumulations of negative ions, giving a polarized ionic distribution in the rock. The effect is most pronounced in rocks that contain clay minerals; first, because the grain and pore sizes are small, and second, because clay grains are relatively strongly charged and adsorb ions on their surfaces. The ionic build-up takes a short time after the voltage is switched on; when the current is switched off, the ions drift back to their original positions.

Electrode polarization is a similar effect that occurs when ore minerals are present. The metallic grains conduct charge by electronic conduction, while electrolytic conduction takes place around them. However, the flow of electrons through the metal is much faster than the flow of ions in the electrolyte, so opposite charges accumulate on facing surfaces of a metallic grain and block the path of ionic flow through the pore fluid. An *overvoltage* builds up for some time after the external current is switched on. The size of the effect is commensurate with the metallic concentration. After the current is switched off, the accumulated ions disperse and the overvoltage decays slowly.

The two effects responsible for induced polarization are indistinguishable at measurement level. The field method for an induced polarization (IP) survey is most often based on the double-dipole array. The current electrodes form a transmitter pair, while the potential electrodes form a receiver pair. The steady-state voltage V_0 is recorded and compared with the amplitude of the decaying residual voltage $V(t)$ at time t after the current is interrupted (Fig. 10.17c). The ratio $V(t)/V_0$ is expressed as a percentage, which decays during the 0.1–10 s between switching the current on and off. If the decay curve is sampled at many points, its shape and the area under the curve may be obtained (Fig. 10.17d). The area under the decay curve, expressed as a fraction of the steady-state voltage, is called the *chargeability* M, defined as

$$M = \frac{1}{V_0}\int_{t_1}^{t_2} V(t)\mathrm{d}t. \tag{10.34}$$

M has the dimensions of time and is expressed in seconds or milliseconds. It is the most commonly used parameter in IP studies.

The induced polarization determines the length of the potential decay time. If it is shorter than the time when the inducing current is off, successive half-cycles of the potential will not interfere. However, if a disseminated conductor is present, the decay time increases, causing overlap and distortion of the half-cycles. The higher the signal frequency the more pronounced is the effect. It increases the ratio $V(t)/V_0$, giving the impression of a better conductor than is really present (i.e., the apparent resistivity decreases with increasing frequency). Clearly, IP and resistivity surveys with alternating current are also influenced. The frequency dependence of the IP effect is exploited by measuring apparent resistivity at two low frequencies. Let these be f and F ($>f$). Commonly $f \approx 0.05$–0.5 Hz and $F \approx 1$–10 Hz. Then $\rho_f > \rho_F$ and we can define a *frequency effect* as

$$\mathrm{FE} = \frac{\rho_f - \rho_F}{\rho_F}. \tag{10.35}$$

The ratio FE is often multiplied by 100 to express it as a percentage (PFE). If no IP effect is present, the resistivity will be the same at both frequencies. The larger the value of FE or PFE, the greater is the induced polarization in the ground. At frequencies above 10 Hz mutual inductance effects between the cables of the primary and detection circuits can produce troublesome potentials, which must be avoided by the field procedure (such as restricting *F*) or minimized analytically.

The presence of metallic conductors is expressed by a similar parameter to FE, the metallic factor (MF). This is proportional to the difference in conductivities at the two measurement frequencies:

$$\begin{aligned}\mathrm{MF} &= A(\sigma_F - \sigma_f) = A(\rho_F^{-1} - \rho_f^{-1}) \\ &= A\left(\frac{\rho_f - \rho_F}{\rho_f \rho_F}\right).\end{aligned} \tag{10.36}$$

The constant *A* is equal to 2π × 10^5; the units of MF are those of conductivity (i.e., $\Omega^{-1}\,\mathrm{m}^{-1}$ or S m^{-1}).

An IP survey includes both lateral profiling and vertical sounding with the expanding spread method. Using a double-dipole array, the distance between nearest electrodes of the transmitter and receiver pairs is a multiple (*na*) of the electrode spacing *a* in each pair. Measurements are made at several discrete positions as the receiver pair is moved incrementally away from the fixed transmitter pair. The transmitter pair is then moved by one increment along the profile and the procedure is repeated. The value of ρ_a, FE, or MF obtained in each measurement is plotted below the mid-point of the array at the intersection of two lines inclined at 45 degrees (Fig. 10.18a). Information is obtained from increasingly greater depths as the transmitter–receiver array expands (i.e., as *n* increases). The plotted value is, however, not the real value of the parameter at the indicated depth. (Recall, for example, that a measurement of apparent resistivity represents an equivalent half-space beneath the array.) A two-dimensional picture of the variation of the IP parameter beneath the profile is synthesized by contouring the results (Fig. 10.18b). The plot is called a *pseudo-section*; it provides a convenient (though artificial) image of the presence of anomalous conductors, but does not represent their true lateral or vertical extent. The presence of anomalous regions may be investigated further by exploratory drilling.

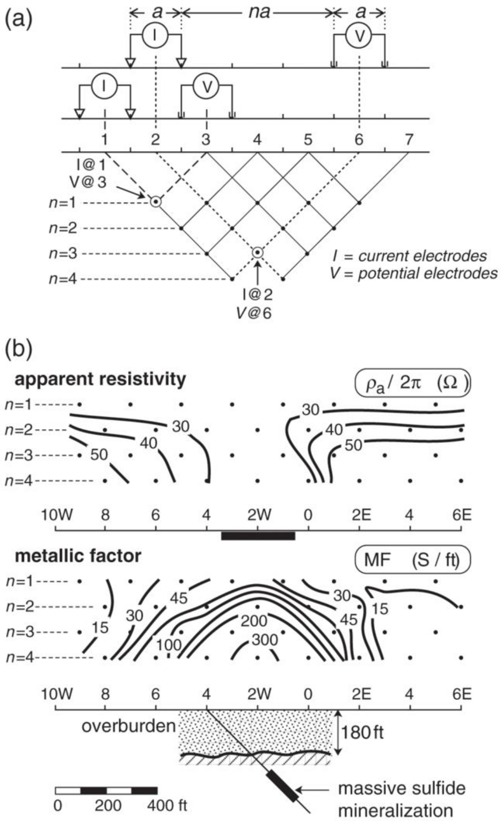

(a)
a
na
a
I
V
I
V
1
2
3
4
5
6
7
I@1
V@3
n=1
n=2
n=3
n=4
I = current electrodes
V = potential electrodes
I@2
V@6
(b)
apparent resistivity
ρa / 2π (Ω)
n=1
n=2
n=3
n=4
30
40
50
30
40
50
10W
8
6
4
2W
0
2E
4
6E
metallic factor
MF (S / ft)
n=1
n=2
n=3
n=4
15
30
45
100
200
300
30
45
15
10W
8
6
4
2W
0
2E
4
6E
overburden
180 ft
massive sulfide
mineralization
0
200
400 ft

Fig. 10.18 (a) Construction of a pseudo-section for a double-dipole IP survey: the measured parameter is plotted at the intersection of 45-degree lines extending from the mid-points of the transmitter and receiver pairs. (b) Pseudo-sections of apparent resistivity and metallic factor for an IP survey over a sulfide orebod.

Source: W.M. Telford, L.P. Geldart and R.E. Sheriff, *Applied Geophysics*, 770 pp., Cambridge: Cambridge University Press, 1990.

Resistivity anomalies depend on the presence of continuous conductors, such as groundwater or massive orebodies. If mineralization is disseminated through a rock it may not cause a significant resistivity anomaly. The good response of the IP method for disseminated concentrations of conducting ore minerals led to its development in base-metal exploration, where large low-grade orebodies may be commercially important. However, the IP effect also depends on the porosity and saturation of the rock. As a result, it can also be used in the search for groundwater and in other environmental applications.

10.5.8 Electrical Resistivity Tomography

The availability of fast, inexpensive computers and the formulation of efficient algorithms has led to the development of electrical tomographic methods akin to the technique of seismic tomography described in Section 7.4.3.2. Seismic tomography based upon teleseismic arrivals from earthquakes is used to describe regions deep in the Earth's mantle that have anomalous seismic velocities. Seismic tomography using refracted and reflected signals from controlled sources can likewise be used to describe velocity perturbations due to shallow features in the Earth's crust. In a similar way, electrical tomography is used to describe the resistivity structure of near-surface regions to depths of several tens of meters. In seismic tomography the observed travel times are inverted to obtain the velocity structure along the path of a seismic ray. Analogously, in resistivity tomography, an inversion procedure is applied to the electrical potentials measured between electrode pairs to obtain the resistivity structure along the current flow lines.

The methods of direct-current resistivity and induced polarization are readily adapted to tomographic analysis. Instead of deploying a single pair of current electrodes and a single pair of potential electrodes, an array of regularly spaced electrodes is deployed. For two-dimensional surveys a linear arrangement of electrodes is used; for three-dimensional investigations the electrodes form an areal array. Various combinations of current electrode pairs and potential electrode pairs are analyzed. The inversion computation is both complex and computation-intensive. It yields a two- or three-dimensional vertical cross-section of the *true* resistivities beneath the electrode array. As in standard resistivity methods, the resolution and maximum depth of investigation depend on the separation and geometry of the electrodes.

The application of direct-current resistivity tomography to an environmental problem is illustrated by an investigation of the extent of ice and permafrost in a buried Alpine rock glacier. The Murtel rock glacier on Mt. Corvatsch in Switzerland is a creeping, permanently frozen (permafrost) body. Its vertical structure is known from a drillhole through the body down to about 50 m depth below the surface. A resistivity tomographic survey employing more than 30 electrodes in Wenner configurations gave a vertical profile of resistivity in good agreement with the drillhole results, and described the lateral extent of the subsurface structure of the permafrost body in detail (Fig. 10.19). Ice has a much higher resistivity than sand or gravel. High resistivities of about 2 MΩm, corresponding to massive ice, are found between 5 m and 15 m depths in the rock glacier. Above and below the ice body, the resistivities are lower. The surface layer resistivity of ~10 kΩm is two orders of magnitude less than in the ice. The region below the ice body is interpreted to consist of frozen sand containing about 30% ice. Its resistivity is an order of magnitude lower than in the ice, but the lower boundary of the ice-block is not clearly defined. Resistivities are less than 5 kΩm in front of the rock glacier, marking the sharp transition to permafrost-free material.

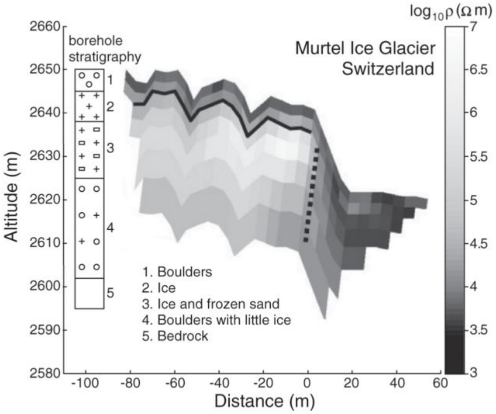


Fig. 10.19 Electrical resistivity tomogram of the Murtel rock glacier, Mt. Corvatsch, Switzerland. The column on the left shows the vertical structure obtained from a borehole at the top of the resistivity profile. Solid lines delineate the bounds of the highly resistive, massive ice body. Dashed lines indicate the interpreted

vertical boundary between the ice body and the permafrost-free material ahead of the glacier.

Source: C. Hauck and D. Von der Muehll, Inversion and interpretation of two-dimensional geoelectrical measurements for detecting permafrost in mountainous regions, *Permafrost Periglac. Process.*, 14, 305–318, 2003.

10.6 Electromagnetic Surveying

The pioneering observations of electrical and magnetic phenomena early in the nineteenth century by Coulomb, Oersted, Ampère, Gauss, and Faraday were unified in 1873 by the Scottish mathematical physicist James Clerk Maxwell (1831–1879). His achievement is of similar stature to that of Newton in gravitation or Einstein in relativity. Like Newton, Maxwell gathered existing knowledge and unified it in a way that allowed the prediction of other phenomena. His book *A Treatise on Electricity and Magnetism* was as important as Newton's *Principia* to the further development of physics. Just as all discussions in dynamics start with Newton's laws, all arguments in electromagnetism begin with Maxwell's equations. In particular, he proposed the theory of the *electromagnetic field,* which classifies light as an electromagnetic phenomenon in the same sense as electricity and magnetism. This ultimately led to the recognition of the wave nature of matter. Unfortunately, Maxwell died while still in the prime of his career, before his theoretical predictions were verified. The German physicist Heinrich Hertz established the existence of electromagnetic waves experimentally in 1887, eight years after Maxwell's untimely death.

Coulomb's law shows that an electric charge is surrounded by an electric field, which exerts forces on other charges, causing them to move, if they are free to do so. Ampère's law shows that an electric charge (or current) moving in a conductor produces a magnetic field proportional to the speed of the charge. If the electric field increases, so that the charge is accelerated, its changing velocity produces a changing magnetic field, which in turn induces another electric field in the conductor (Faraday's law) and thereby influences the movement of the accelerated charge. The coupling of the electric and magnetic fields is called electromagnetism. If two straight conductors are laid end-to-end and connected in series, they act as an electrical dipole. An alternating electric field applied to the conductors causes the dipole to oscillate, acting as an antenna for the emission of an electromagnetic wave. This consists of a magnetic field *B* and an electric field *E*, which vary with the frequency of the oscillator, and are oriented at right angles to each other in the plane perpendicular to the direction of propagation (Fig. 10.20). In a vacuum all electromagnetic waves travel at the speed of light ($c = 2.99792458 \times 10^8$ m s^{-1}; about 300,000 km s^{-1}), which is one of the fundamental constants of nature.

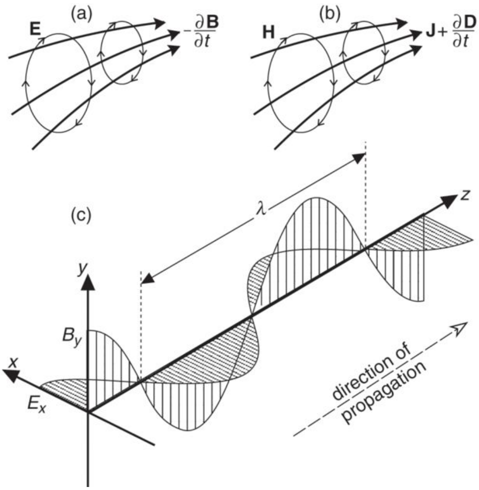

(a)
E
$-\frac{\partial \mathbf{B}}{\partial t}$
(b)
H
$\mathbf{J}+\frac{\partial \mathbf{D}}{\partial t}$
(c)
λ
z
y
B_y
x
E_x
direction of propagation

Fig. 10.20 (a) An electric field **E** is generated by a changing magnetic field ($\partial \mathbf{B}/\partial t$), while (b) a magnetic field **B** is produced by the current density **J** and the changing displacement-current density ($\partial \mathbf{D}/\partial t$); (c) in an electromagnetic wave an electric field E_x and a magnetic field B_y fluctuate normal to each other in the plane normal to the propagation direction (*z*-axis).

The derivation of electromagnetic field equations from Maxwell's equations is beyond the level of this textbook, but their meaning can be readily understood. Two equations, identical in form, are obtained. They describe the propagation of the **B** and **E** field vectors, respectively, and are written as:

$$\nabla^2 \mathbf{B} = \mu_r \mu_0 \sigma \frac{\partial \mathbf{B}}{\partial t} + \mu_r \mu_0 \varepsilon_r \varepsilon_0 \frac{\partial^2 \mathrm{B}}{\partial t^2},$$

(10.37)

$$\nabla^2\mathbf{E} = \mu_r\mu_0\sigma\frac{\partial \mathbf{E}}{\partial t} + \mu_r\mu_0\varepsilon_r\varepsilon_0\frac{\partial^2 \mathbf{E}}{\partial t^2},$$

(10.38)

where, in Cartesian coordinates, $$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

In these equations σ is the *electrical conductivity;* μ_r is the *magnetic permeability*, which in most materials (unless they are ferromagnetic) is very close to 1; μ_0 is the *magnetic constant* (Box 11.1), or permeability of free space ($\mu_0 = 4\pi \times 10^{-7}$ N A^{-2}); ε_r is the *relative permittivity* of the material; and ε_0 is the permittivity constant, or permittivity of free space ($\varepsilon_0 = 8.854187817 \times 10^{-12}$ C^2 N^{-1} m^{-2}). The quantities μ_0 and ε_0 are related to the speed of light c by the equation $\mu_0\varepsilon_0 = 1/c^2$. The SI definitions of μ_0 and c are exact, thus ε_0 is also defined exactly. In constant electrical fields ε_r is known as the *dielectric constant*, κ. The value of κ is ~ 5–20 in most rocks and minerals and 80 in water (Section 10.2.3). In sediments and sedimentary rocks, the water content plays an important role in determining the value of κ. The value of ε_r increases with the frequency of the electrical field.

Electromagnetic radiation encompasses a wide frequency spectrum. It extends from very high-frequency (short-wavelength) γ-rays and X-rays to low-frequency (long-wavelength) radio signals (Fig. 10.21). Visible light constitutes a narrow part of the spectrum. Two ranges of electromagnetic radiation are of particular importance in solid Earth geophysics: a high-frequency range in the radar part of the spectrum, and a broad range of low frequencies extending from audio frequencies to signals with periods of hours, days, or years. The electromagnetic equations reduce to simpler forms for these two particular frequency ranges.

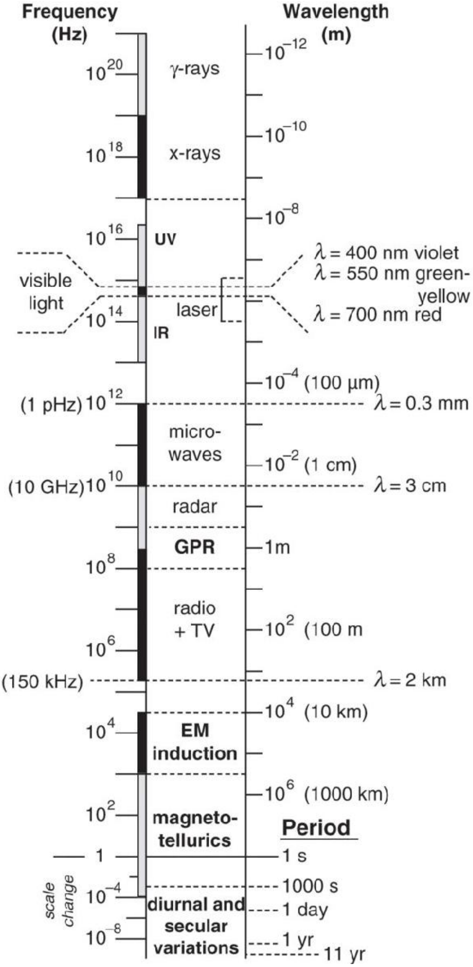

Frequency (Hz)
Wavelength (m)
10^{20}
γ-rays
10^{-12}
10^{18}
x-rays
10^{-10}
10^{16}
UV
10^{-8}
λ = 400 nm violet
λ = 550 nm green-yellow
λ = 700 nm red
visible light
laser
10^{14}
IR
10^{-4} (100 μm)
(1 pHz) 10^{12}
λ = 0.3 mm
micro-waves
10^{-2} (1 cm)
(10 GHz) 10^{10}
λ = 3 cm
radar
GPR
1m
10^{8}
radio + TV
10^{2} (100 m
10^{6}
(150 kHz)
λ = 2 km
10^{4} (10 km)
10^{4}
EM induction
10^{6} (1000 km)
10^{2}
magneto-tellurics
Period
1
1 s
scale change
10^{-4}
1000 s
diurnal and secular variations
1 day
10^{-8}
1 yr
11 yr

Fig. 10.21 The electromagnetic spectrum, showing the frequency and wavelength ranges of some common phenomena and the frequencies and periods used in electromagnetic surveying.

The left side of Eq. (10.38) describes the variation of the *E*-component of the electromagnetic wave in space. The right side describes its variation with time and so its frequency dependence. The first term is related to the familiar *conduction* of electricity in a conductor. Maxwell introduced the second term and called it the *displacement* current. It originates when charges are displaced but not separated from their atoms, causing an electric polarization; fluctuations in the polarization have the effect of an alternating displacement current. Suppose that the electric dipole emitting *E* and

B oscillates sinusoidally with angular frequency ω. Then $|\partial E/\partial t| \sim \omega E$, and $|\partial^2 E/\partial t^2| \sim \omega^2 E$, so the *magnitude ratio* (MR) of the second (*displacement*) term to the first (*conduction*) term on the right side of Eq. (10.38) is

$$\mathrm{MR} = \frac{\left|\mu_0 \varepsilon_0 \varepsilon_\mathrm{r} \dfrac{\partial^2 E}{\partial t^2}\right|}{\left|\mu_0 \sigma \dfrac{\partial E}{\partial t}\right|} = \frac{\varepsilon_0 \varepsilon_\mathrm{r} \omega^2 E}{\sigma \omega E} = 2\pi f \frac{\varepsilon_0 \varepsilon_\mathrm{r}}{\sigma}, \tag{10.39}$$

where f is the frequency of the signal. The conductivity σ of rocks and soils is generally in the range 10^{-5} to 10^{-1} S m^{-1}; in orebodies σ may be as large as 10^3–10^5 S m^{-1} (see Fig. 10.3). Electromagnetic induction surveying is usually carried out at frequencies below 10^5 Hz, for which the magnitude ratio MR is much less than unity in both good and bad conductors. At these frequencies the electromagnetic signal passes through the ground in a diffusive manner, by conversion of the changing magnetic fields to electric currents and vice versa. High-frequency surveying employs radar signals with frequencies around 10^8 Hz, for which the magnitude ratio MR is very small in an orebody but can be much greater than unity in rocks and soils. Under these conditions the electromagnetic signal propagates like a wave, and so is subject to diffraction, refraction, and reflection.

10.6.1 Electromagnetic Induction

Electromagnetic (EM) surveys carried out at frequencies below 100 kHz are based on Faraday’s law of electromagnetic induction. An alternating magnetic field in a coil or cable induces electric currents in a conductor. The conductivity of rocks and soils is poor and currents induced in them are weak compared to the induction effect in a good conductor. *Eddy currents* induced in the ground produce secondary magnetic fields. These are superposed on the primary field and are measured at the ground surface (Fig. 10.22a).

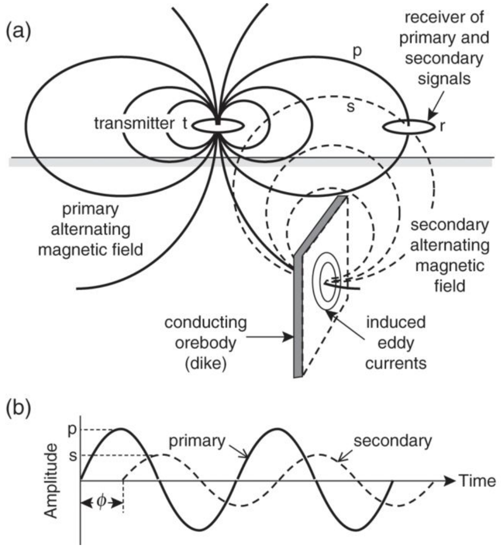

(a)
receiver of primary and secondary signals
p
s
r
transmitter t
primary alternating magnetic field
secondary alternating magnetic field
conducting orebody (dike)
induced eddy currents
(b)
Amplitude
p
s
primary
secondary
Time
ϕ

Fig. 10.22 (a) Illustration of primary and secondary fields in the horizontal loop induction method of electromagnetic exploration for shallow orebodies. (b) Amplitudes and phases of the primary (p) and secondary (s) fields.

Suppose that a low-frequency plane wave propagates along the vertical z-axis. At low frequencies the displacement current is negligible compared to the conduction current and Eq. (10.37) becomes

$$\frac{\partial^2 \mathbf{B}}{\partial z^2} = \mu_0 \sigma \frac{\partial \mathbf{B}}{\partial t}, \tag{10.40}$$

where the magnetic field has components B_x and B_y. The form of this equation is reminiscent of the one-dimensional equation of *diffusion* or *heat conduction* (Eq. 9.28), whose solution (Eq. 9.30) describes how the temperature changes with time and position when a fluctuating temperature acts on the surface. By analogy, the solution of Eq. (10.40) for the components B_x or B_y of an alternating magnetic field with angular frequency ω ($= 2\pi f$) in a conductor with conductivity σ is

$$B_{x,y}(z,t) = B_0 e^{-z/d} \cos\left(\omega t - \frac{z}{d}\right), \tag{10.41}$$

$$\text{where} \qquad d = \sqrt{\frac{2}{\mu_0 \sigma \omega}}. \tag{10.42}$$

Here, d is called the *skin depth*. At this depth the magnetic field is attenuated to e^{-1} (~37%) of its value outside the conductor. The skin depth is dependent on the conductivity of the body and the frequency of the field. The skin depth in normal ground ($\sigma \sim 10^{-3}$ S m^{-1}) for a low-frequency alternating magnetic field ($f \sim 10^3$ Hz) is about 500 m, but in an orebody ($\sigma \sim 10^4$ S m^{-1}) it is only ~16 cm. The comparable figures for a high-frequency radar signal ($f \sim 10^9$ Hz) are 50 cm and 0.16 mm, respectively. Note that the skin depth is not the maximum depth of penetration of the magnetic field. It helps to indicate how rapidly the field is attenuated, but the magnetic field is effective at depths that are many times the skin depth. However, it decays to ~1% at a depth $z = 5d$ and to ~0.1% at $z = 7d$, effectively limiting the practical depth of exploration with the induction method.

The many field methods of EM induction have a common principle. A coil or cable is used as the transmitter of the primary alternating magnetic field, while another coil serves as the receiver of both the primary signal and a secondary signal from the eddy currents induced in a conductor (Fig. 10.22a). The magnetic field in the conductor experiences a phase shift (equal to z/d; Eq. 10.41) due to the conductivity. This results in a phase difference ϕ between the secondary and primary signals in the receiver (Fig. 10.22b). The exact theory of EM induction is complicated, even in a simple situation, but we can obtain a simple qualitative appreciation by applying some concepts from electrical circuit theory. As in Box 9.2, we will use complex numbers involving i, the square root of –1 (Box 4.6).

Let the current systems in transmitter, receiver, and conductor be represented by simple loops carrying currents I_t, I_r, and I_c, respectively. If the currents are sinusoidal, each has the form $I = I_0 e^{i\omega t}$, so that $dI/dt = i\omega I$. Let the resistance of the conductor be R and its self-inductance be L. The voltage V_c in the conductor is composed of two parts. A resistive part due to the current I_c in the resistance R is equal to I_cR. An inductive part due to the change of current is equal to $L\, dI_c/dt$. The complete voltage in the conductor is then

$$V_c = I_cR + L\frac{dI_c}{dt} = I_c(R + i\omega L). \tag{10.43}$$

The voltage V_c is induced in the conductor by the changing current I_t in the transmitter circuit. Let the mutual inductance between transmitter and conductor be M_{tc}; then $V_c = -M_{tc}\, dI_t/dt$. Similarly, the transmitter current induces a primary voltage $V_p = -M_{tr}\, dI_t/dt$ in the receiver, in which the eddy currents in the conductor also induce a secondary voltage $V_s = -M_{cr}\, dI_c/dt$. Here, M_{tr} and M_{cr} are the mutual inductances between transmitter and receiver, and conductor and receiver, respectively. The following relationships exist between the different voltages and currents:

$$V_{\mathrm{p}} = -M_{\mathrm{tr}}\frac{dI_{\mathrm{t}}}{dt} = -\mathrm{i}\omega M_{\mathrm{tr}}I_{\mathrm{t}}, \tag{10.44}$$

$$V_{\mathrm{s}} = -M_{\mathrm{cr}}\frac{dI_{\mathrm{c}}}{dt} = -\mathrm{i}\omega M_{\mathrm{cr}}I_{\mathrm{c}}, \tag{10.45}$$

$$V_{\mathrm{c}} = -M_{\mathrm{tc}}\frac{dI_{\mathrm{t}}}{dt} = -\mathrm{i}\omega M_{\mathrm{tc}}I_{\mathrm{t}}. \tag{10.46}$$

Combining Eqs. (10.43) and (10.46), we get

$$\frac{I_{\mathrm{c}}}{I_{\mathrm{t}}} = \frac{-\mathrm{i}\omega M_{\mathrm{tc}}}{R + i\omega L} = \frac{-\mathrm{i}\omega M_{\mathrm{tc}}}{(R^2 + \omega^2 L^2)}(R - \mathrm{i}\omega L). \tag{10.47}$$

From Eqs. (10.44) and (10.45) the ratio of V_{s} to V_{p} in the receiver is

$$\frac{V_{\mathrm{s}}}{V_{\mathrm{p}}} = \frac{M_{\mathrm{cr}}}{M_{\mathrm{tr}}}\frac{I_{\mathrm{c}}}{I_{\mathrm{t}}} = -\frac{M_{\mathrm{tc}}M_{\mathrm{cr}}}{M_{\mathrm{tr}}}\frac{(\omega^2 L + \mathrm{i}\omega R)}{(R^2 + \omega^2 L^2)}, \tag{10.48}$$

which can be written in the form

$$\frac{V_{\mathrm{s}}}{V_{\mathrm{p}}} = \frac{M_{\mathrm{cr}}}{M_{\mathrm{tr}}}\frac{I_{\mathrm{c}}}{I_{\mathrm{t}}} = -\frac{M_{\mathrm{tc}}M_{\mathrm{cr}}}{M_{\mathrm{tr}}L}\left(\frac{\beta^2 + i\beta}{1 + \beta^2}\right), \tag{10.49}$$

where $\beta = \omega L/R$ is the *response parameter* of the conductor. The function in parentheses in Eq. (10.49) is a complex number, so the voltage ratio (or *response* of the measuring system) can be written $P + iQ$. The real part P has the same phase as the primary signal and is called the *in-phase component* of the response. The imaginary part Q is 90° out of phase with the primary signal (i.e., if the primary signal is $\sim\cos \omega t$, the imaginary part is $\sim\sin \omega t = \cos [\omega t - \pi/2]$); it is called the *quadrature component*.

10.6.2 EM Induction Surveying

The resemblance of the basic equations of EM induction to the diffusion equation (Eq. 9.29) classifies the method as a *diffusive* one. Diffusive techniques – for example, the gravity, magnetic field, geothermal, and seismic surface wave methods – respond to a volumetric average of the specific physical parameter and do not show fine detail of its distribution. Hence, EM induction yields an average value of the electrical conductivity in a particular volume, but the resolution is better than that of potential field methods.

In the simplest ground-based EM induction method the transmitter may be a long cable or large horizontal loop on the ground surface, or it may be a small coil (diameter ~1 m) with its axis vertical or horizontal. The receiver is usually a similar small coil. It can be used to detect the direction, intensity, or phase of the secondary signal. In its most simple application the tilt of the coil about a horizontal axis measures the dip-angle of the combined primary and secondary fields at the receiver. The method allows location, outlining and, to some extent, depth determination of a conductor. However, the dip-angle method registers only part of the available information in the secondary signal and does not describe the electrical properties of the conductor. As shown by Eq. (10.49), these properties affect the phase and relative amplitudes of the in-phase and quadrature components of the secondary signal relative to the primary. Phase-component EM measurement methods, therefore, allow more detailed interpretation.

Ground-based surveying is exemplified by the horizontal loop electromagnetic method (HLEM), popularly known also by the commercial names *Slingram* or *Ronka*. The receiver and transmitter are coupled by a fixed cable about 30–100 m in length, and kept at a constant separation while the pair is moved along a traverse of a suspected conductor (Fig. 10.23a). The cable supplies a direct signal that exactly cancels the primary signal at the receiver, leaving only the secondary field of the conductor. This is separated into in-phase and quadrature components, which are expressed as percentages of the primary field and plotted against the position of the mid-point of the pair of coils (Fig. 10.23b). The in-phase and quadrature signals are zero far from the conductor and at the places where either the transmitter or receiver passes over the conductor. This enables the outline of a buried conductor to be charted. The signal rises to a positive peak on either side and falls to a negative peak over the middle of the conductor. The peak-to-peak responses of the in-phase and quadrature components depend on the quality of the conductor, which is expressed by a response parameter such as β in Eq. (10.49). A suitable function is the dimensionless parameter $\alpha = \mu_0\sigma\omega\, sl$, which contains the conductivity σ and width s of the conductor as well as the coil spacing l and frequency ω of the EM system. The systematic variation of the response curves with the value of α (Fig. 10.23b) allows interpretation of the quality of the conductor. A simple way of doing this is with the aid of model response curves. The variation of in-phase and quadrature signals over a conducting orebody can be modeled experimentally on a smaller scale in the laboratory. The smaller values of s and l are compensated by larger values of σ and ω to give the same response parameter α. The model response curves for different α are then directly applicable to the interpretation of real conductors measured in the field.

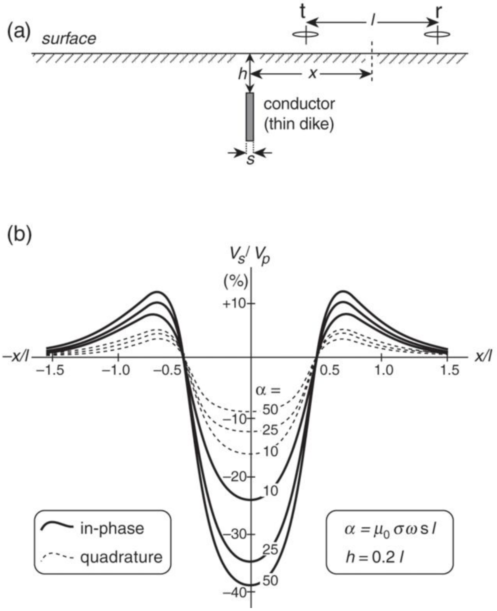


Fig. 10.23 (a) Geometry of an HLEM profile across a thin vertical dike. (b) In-phase and quadrature profiles over a dike at depth $h/l = 0.2$ for some values of the dimensionless response parameter α.

The most common procedure in EM induction surveys is by measuring parallel profiles, usually on traverses at right angles to the geological strike of dikes or other suspected conducting bodies. In environmental investigations EM induction is useful for locating buried pipes that may carry fluids or gases. Ground-based EM methods may also be used for vertical sounding, applying the same principles as in resistivity methods to obtain the conductivities of horizontal layers. The greater the separation of transmitter and receiver, the deeper is the maximum depth at which conductors may be analyzed. The depth penetration also depends on the frequency of the EM signal: low frequencies penetrate more deeply than high frequencies. As a result, higher frequencies are more useful in environmental investigations of shallow features, such as groundwater depths, whereas a mineralization of interest to mining companies may be more deeply buried and require exploration at lower frequencies.

The EM induction method is very suitable for airborne surveys.

These were carried out originally with fixed wing aircraft but they now more commonly use helicopters, which can adapt better to terrain roughness while flying close to the ground. The transmitter and receiver coils are mounted for optimum coupling (e.g., with their axes coaxial and parallel to the line of flight) in fixed positions in the aircraft or in a towed "bird," using similar configurations as in airborne magnetometer surveys (see Fig. 11.24a). Alternatively, the transmitter may be in the airplane and the receiver in the bird or in another airplane. The increased separation of transmitter and receiver in this configuration gives greater depth penetration. However, in flight the bird yaws and pitches, altering the separation and parallelism of the coils, so that often only the quadrature component is usable. The flight patterns consist of parallel profiles traversing the terrain. Lines are flown at about 100 m above ground level with fixed-wing aircraft and 30 m with helicopters.

In order to expand the range of subsurface depths that can be explored simultaneously, modern EM induction systems operate at several frequencies. A multi-frequency coil arrangement for airborne exploration, housed in a 10 m long towed "bird" for a helicopter survey, is sketched in Fig. 10.24. In this example, the bird contains five pairs of coplanar horizontal coils and a single coaxial pair, operated at single frequencies from 400 Hz to 140 kHz. The separation between transmitter and receiver is around 8 m for each coplanar pair and 9 m for the coaxial pair. Different coil configurations and operating frequencies are used as required by the parameters and goals of an investigation. The location of the bird is determined precisely by differential GPS (Section 3.5.4) and the altitude above ground is measured by a laser altimeter.

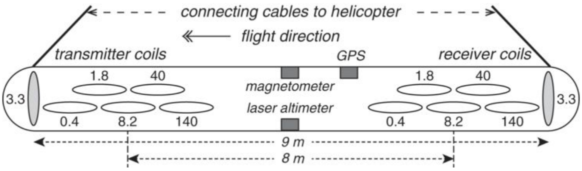


Fig. 10.24 Sketch (not to scale) of the RESOLVE helicopter-borne

EM system consisting of five pairs of horizontal coplanar coils and a single pair of coaxial coils, towed in a cylindrical housing beneath a helicopter. Numbers beside the coils are operating frequencies in kHz.

Source: EM GeoSci website, https://em.geosci.xyz/index.html.

10.6.3 Transient Electromagnetic Surveying

Electromagnetic induction methods that determine the presence and quality of a conductor by measuring the in-phase and quadrature components of the secondary magnetic field dominated electromagnetic prospecting until the 1980s and are still successfully employed. Advances in electronics and a huge increase in inexpensive computational power made the transient electromagnetic (TEM) method increasingly important. This is a computationally intensive method for use in both ground-based and airborne investigations. Initially installed as a large transmitting coil around a fixed-wing aircraft towing a receiving coil, it is now common to use a slower but more maneuverable helicopter (Fig. 10.25a). Instead of transmitting discrete frequencies, the TEM method uses a pulsed direct current as source. The waveform has an almost square shape with a sharp cutoff; this induces eddy currents with numerous frequencies which decay with time at different rates depending on the ground conductivity.

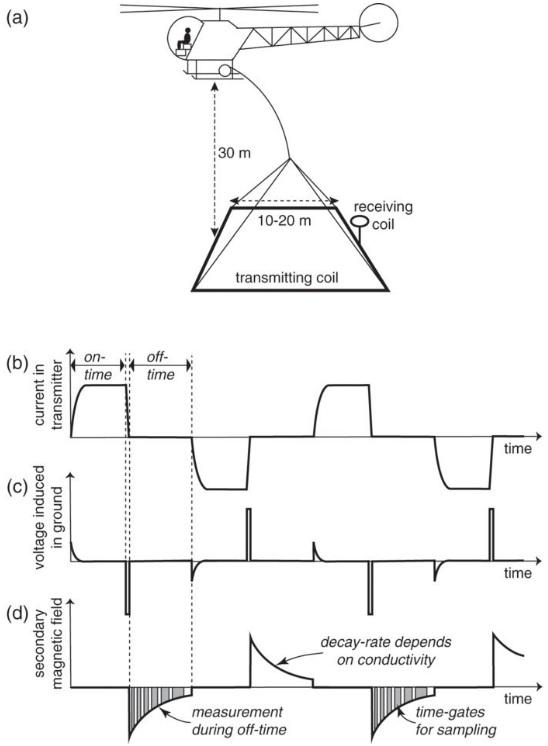

(a)
30 m
10-20 m
receiving coil
transmitting coil
(b)
current in transmitter
on-time
off-time
time
(c)
voltage induced in ground
time
(d)
secondary magnetic field
decay-rate depends on conductivity
time
measurement during off-time
time-gates for sampling

Fig. 10.25 (a) Schematic drawing of airborne transient electromagnetic surveying with a helicopter. (b) Current in the primary, transmitting coil. (c) Electromotive force induced in the ground by the primary field of the transmitting coil. (d) Secondary magnetic field detected by the receiver.

Source: A. V. Christensen, E. Auken, and K. Srensen. The transient electromagnetic method. In: *Groundwater Geophysics*, R. Kirsch, ed., pp. 179–226, Berlin: Springer, 2009.

In the TEM method a direct current in a transmitting coil produces a steady primary magnetic field. The field is maintained at a constant value for 10–20 ms, then switched off abruptly, e.g., within a millisecond (Fig. 10.25b). An electromotive force (emf) is induced in the ground, proportional to the rate of change of flux (Fig. 10.25c), which is made as rapid as possible. As in other EM methods an electrical current is induced in the ground and produces a secondary magnetic field that is detected by the receiving system. The primary current is off for 10–20 ms, during which time the current in the ground decays (Fig. 10.25d), at a rapid rate in a poor conductor and slowly in a good conductor. The next half-cycle repeats the process with the current in the opposite direction, resulting in a pulsing rate of 10–20 cycles per second.

The primary magnetic field caused by the abrupt shut-off of the transmitter current induces electrical currents with numerous frequencies in the ground. Because of the skin-depth effect, each frequency penetrates to a different maximum depth, depending on the vertical conductivity profile. Therefore the decaying secondary field also consists of the superposition of many frequencies, whose amplitudes have been modified by the ground conductivity. In general, the early part of the decaying secondary field reflects the conductivity in shallow depths, while the later part is due to deeper sources. However, this is an oversimplified model of the field at the receiver, because the secondary magnetic fields of the eddy currents in each depth induce further eddy currents and magnetic fields from deeper conductive layers, which contribute to the decay curve.

The decay curve is sampled during a number of predetermined short time-windows (called gates) during the off-time of the transmitter to minimize interference from the primary field. The sampling gates have variable widths. They are very short in the early part of the decaying interval and longer in the later part where the secondary field is weaker. The gated sampling is repeated for each pulse and the sampled values for each specific gate are stacked (added together) in order to enhance the very weak secondary signals. The computational technique of Fourier analysis is used to analyze the gated decay curve, from which the subsurface conductivity profile is interpreted.

In ground-based operation of the TEM method, a square coil is laid on the ground as the transmitter and a small coil at the middle of the layout serves as the receiver. The edge length of the transmitter is chosen for optimum depth analysis. It may be only a few meters for a near-surface environmental study, or as much as 100 m for deeper investigation. In airborne exploration with a helicopter (Fig. 10.25a), the transmitting coil may be 40 m or more in diameter, suspended 30–100 m below the aircraft. The flight altitude increases the distance to subsurface conducting bodies and thus reduces the possible resolution compared to a ground-based method, but this is compensated by the speed and ease of usage over inaccessible regions. The method is used for mineral exploration as well as environmental investigations.

10.6.4 Magnetotelluric Exploration

Magnetotelluric (MT) exploration (also referred to as MT sounding) is a natural-source electromagnetic method. It is an important form of vertical EM sounding that takes advantage of the ability of low-frequency signals from natural sources in the external geomagnetic field to penetrate deeply into the crust and mantle. The fluctuating electromagnetic fields that originate in the ionosphere are partly reflected at the Earth's surface; the returning fields are again reflected off the conducting ionosphere. This happens repeatedly, so that the fields eventually have a strong vertical component and may be regarded as vertically propagating plane waves with a wide spectrum of frequencies. These fields penetrate into the ground and induce telluric electric currents (Section 10.4.3), which in turn generate secondary magnetic fields. The telluric currents are detected with two pairs of electrodes, usually oriented north–south and east–west. Three components of the magnetic fields are measured: the vertical component and a horizontal component parallel to each of the telluric components. The method yields conductivity information from much greater depths than artificial-source induction methods. It has been applied in the search for petroleum and deep zones of mineralization in the upper crust. Utilizing long periods in the range 10–1000 s, it is an important method for the investigation of the structure of the crust and upper mantle.

Consider a plane electromagnetic wave propagating in the z-direction (Fig. 10.20). Let the electric component E_x be along the x-axis, so that the magnetic field B_y (normal to E_x) is along the y-axis. Ampère's law, as summarized in Maxwell's equations, relates E_x to the gradient of B_y in the z-direction. Because B has only a y-component, Ampère's law reduces to

$$E_x = -\frac{1}{\mu_0 \sigma}\frac{\partial B_y}{\partial z},$$

(10.50)

where B_y has the form of Eq. (10.41). Differentiating B_y by parts gives

$$\begin{aligned}E_x &= -\frac{B_0}{\mu_0\sigma}\left\{\left(-\frac{e^{-z/d}}{d}\right)\cos\left(\omega t-\frac{z}{d}\right)\right.\\ &\qquad\left.+e^{-z/d}\left(-\frac{1}{d}\right)\left(-\sin\left(\omega t-\frac{z}{d}\right)\right)\right\}\\ &=\frac{B_0}{\mu_0\sigma d}e^{-z/d}\left\{\cos\left(\omega t-\frac{z}{d}\right)-\sin\left(\omega t-\frac{z}{d}\right)\right\}\\ &=\frac{B_0}{\mu_0\sigma d}e^{-z/d}\sqrt{2}\cos\left(\omega t-\frac{z}{d}+\frac{\pi}{4}\right).\end{aligned}$$

(10.51)

Comparison of Eqs. (10.41) and (10.51) shows a phase shift of 45° (π/4) between E_x and B_y. However, the ratio of the maximum amplitudes of the two components is

$$\frac{|E_x|}{|B_y|}=\frac{\sqrt{2}}{\mu_0\sigma d}.$$

(10.52)

If we now substitute for d from Eq. (10.42) and write $\rho = 1/\sigma$, we get

$$\rho=\frac{\mu_0}{\omega}\frac{|E_x|^2}{|B_y|^2},$$

(10.53)

$$d=\frac{2}{\omega}\frac{|E_x|}{|B_y|},$$

(10.54)

for the effective resistivity ρ at depth d. The analysis gives similar results for an electric field along the y-axis and a corresponding magnetic field along the x-axis. In this case the ratio of the field amplitudes is $|E_y|/|B_x|$.

In addition to the horizontal magnetic fields, B_x and B_y, the vertical component B_z is also recorded for use in the interpretation of two-dimensional structures. Thus the dataset from an MT site consists of two electrical components and three magnetic field components recorded continuously during a lengthy observation interval covering some hours or days. The recorded magnetic fields consist of an external part from the ionosphere and an internal part related to the induced current distribution. These components must be separated analytically. The electric and magnetic records contain numerous frequencies, some of which are simply noise and some of which are of geophysical interest. As a result, sophisticated data-processing is required, involving power-spectrum analysis and filtering.

The interpretation of MT data is based on either modeling or inversion. The modeling method is a direct approach to solving the conductivity distribution. It assumes a conductivity model for which a theoretical response is calculated and compared with the real response. The parameters of the model are adjusted in turn repeatedly to obtain the most favorable fit to the observations. As in the case of vertical electrical sounding with direct currents (Section 10.5.6), the inversion method seeks a solution to the EM induction problem by using the frequency spectrum of the observations to establish the causal conductivity distribution.

Although MT sounding can be carried out in the subaudio to audio range ($f \sim 10\text{–}10^4$ Hz), its main application is in determining the electrical conductivity at great depths using very low frequencies ($f \ll 1$ Hz). The investigation of resistivity in the crust and upper mantle using MT sounding is illustrated by a profile across Vancouver Island (Fig. 10.26). Twenty-seven MT sounding stations were located along a northwest–southeast reflection seismic profile. One-dimensional analysis of the vertical distribution of resistivity beneath three stations (10, 12, and 14) showed an electrical discontinuity at virtually the same depth as a major seismic reflector observed in the associated seismic reflection profile. The information acquired about the mean resistivity above this depth was then used in a two-dimensional inversion of the MT records at all the stations. The resistivity pattern was interpreted down to depths of 100 km. It shows a northeastward dipping zone of low resistivity (ρ ≈ 30 Ω m), referred to as the E-conductor, surrounded by much more resistive material (ρ ≥ 5000 Ω m). The E-conductor was interpreted as the top of the descending Juan de Fuca plate, where it subducts under the North American plate. The anomalously high conductivity in the top of the plate was attributed to conducting fluids in sediments derived from the accretionary wedge.

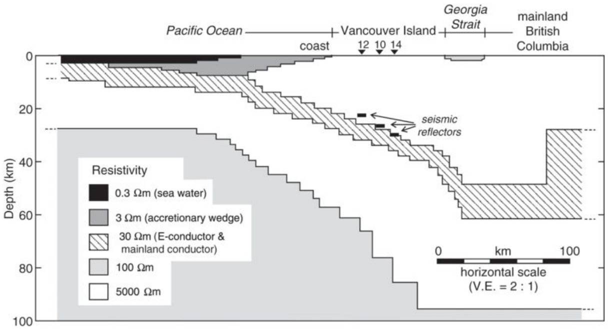

Pacific Ocean
coast
Vancouver Island
12 10 14
Georgia Strait
mainland British Columbia
seismic reflectors
Depth (km)
0
20
40
60
80
100
Resistivity
0.3 Ωm (sea water)
3 Ωm (accretionary wedge)
30 Ωm (E-conductor & mainland conductor)
100 Ωm
5000 Ωm
0
km
100
horizontal scale
(V.E. = 2 : 1)

Fig. 10.26 Two-dimensional resistivity model of the crust and upper mantle beneath Vancouver Island and the adjacent mainland derived from magnetotelluric results.

Source: R.D. Kurtz, J.M. DeLaurier and J.C. Gupta, A magnetotelluric sounding across Vancouver Island detects the subducting Juan de Fuca plate, *Nature* 321, 596–599, 1986.

10.6.5 Ocean Current Circulation and Upper Mantle Conductivity

Seawater is a good conductor of electricity because it has a high concentration of dissolved salts. The oceans are constantly in motion, driven, above all, by tidal forces. The motion of the electrically conducting seawater in the Earth's magnetic field induces weak electrical currents in the oceans as a result of the Lorentz force (Box 11.1). These electrical currents generate secondary magnetic fields with amplitudes of several nanotesla (nT) at the ocean surface, which can be detected by magnetometers stationed on land and especially in satellites. At the height of the satellite orbits these magnetic fields have amplitudes of less than 2 nT, compared to the main geomagnetic field intensity of up to 54,000 nT at that altitude; nevertheless, they can be mapped and interpreted in terms of oceanic current flow.

The ocean surface has topographic highs and lows that reflect the heights of local geoid undulations. Driven by gravity, seawater flows away from a topographic high toward a lower level, which produces ocean currents. The flow of conductive seawater through the ambient geomagnetic field induces electrical currents in the seawater; these currents are the sources of secondary magnetic fields. The CHAMP and SWARM low-orbit satellite magnetic missions (Sections 3.5.5 and 3.5.6) have been able to measure the weak magnetic fields due to the oceanic electrical currents and to separate and identify the component due to the lunar semi-diurnal tides. The flow of seawater in the circulation systems in ocean basins produces magnetic fields that are weaker than the tidal effect; they are difficult to separate from the background of various other electromagnetic signals.

The CHAMP and SWARM magnetic recordings originating from the semi-diurnal oceanic tides and the magnetosphere have been used to estimate the electrical conductivity structure of the mantle. The structure is expressed in vertical depth profiles of electrical conductivity (Fig. 10.27). Inversion of the tidal signals shows an upper resistive layer (conductivity ~2 × 10^{-4} S m^{-1}), estimated to be around 70 km thick, below which the conductivity increases rapidly, and then steadies below 250 km depth at ~1 × 10^{-1} S m^{-1}. The increase can be interpreted as the transition from cold rigid lithosphere to a hotter asthenosphere containing partial melts and water. Inversion of the magnetospheric signals yields a conductivity profile that does not recognize a lithosphere–asthenosphere boundary but increases smoothly through the upper mantle transition zone. Joint inversion of both datasets shows that the mantle conductivity below the transition zone is on the order of 1 S m^{-1}.

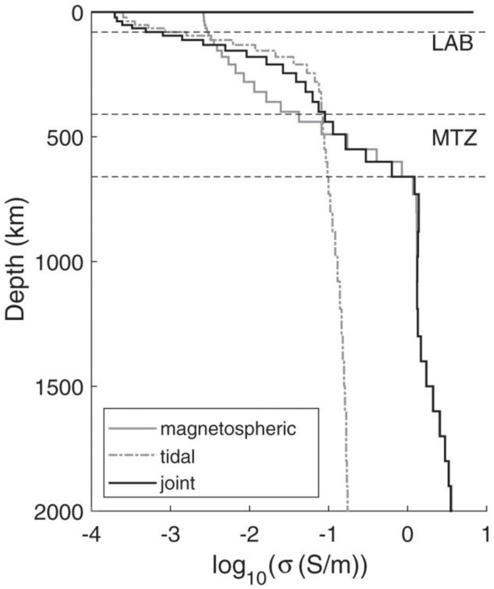


Fig. 10.27 Models of electrical conductivity in the mantle derived from satellite magnetospheric (gray solid), and semidiurnal oceanic tidal (gray dashed) signals, inverted separately and jointly (black solid). LAB is the mean depth of the lithosphere–asthenosphere boundary under the oceans and MTZ is the upper mantle transition zone.

Source: Grayver, A. V., Munch, F. D., Kuvshinov, A., Khan, A., Sabaka, T. J. and Tøffner-Clausen, L. 2017. Joint inversion of satellite-detected tidal and magnetospheric signals constrains electrical conductivity and water content of the upper mantle and

transition zone. *Geophys. Res. Lett.*, 44, 6074–6081, 2017.

10.6.6 Ground Penetrating Radar

At high frequencies in poorly conducting media the conduction term in the electromagnetic equations is negligible compared to the displacement term. The electric field equation then becomes

$$\nabla^2 E = \mu_0 \varepsilon \frac{\partial^2 E}{\partial t^2}, \tag{10.55}$$

with a similar equation for the magnetic field. This has the familiar form of the wave equation, which describes the propagation of an elastic disturbance (Eq. 6.4). Analogously, Eq. (10.55) describes the propagation of the electric part of an electromagnetic wave. By comparing with the seismic wave equations we see that the **E** and **B** fields in an electromagnetic wave have the same velocity V, where $V^2 = 1/\mu_0\varepsilon$. In a vacuum the wave velocity is equal to the velocity of light c, given by $c^2 = 1/\mu_0\varepsilon$. Using the relationship $\varepsilon = \kappa\varepsilon_0$ we get $V^2 = c^2/\kappa$, and taking into account that the dielectric constant κ is ~5–20 in earth materials, the velocity of an electromagnetic wave in the ground is found to be ~ $0.2c$–$0.6c$.

To simplify further discussion, suppose that the electromagnetic disturbance propagates along the z-axis (i.e., $\partial/\partial x = \partial/\partial y = 0$), so that $\nabla^2 = \partial^2/\partial z^2$. For this one-dimensional case,

$$\frac{\partial^2 E}{\partial z^2} = \mu_0 \varepsilon \frac{\partial^2 E}{\partial t^2} = \frac{1}{V^2}\frac{\partial^2 E}{\partial t^2}. \tag{10.56}$$

If we compare this equation with Eq. (6.4) and its solution (Eq. 6.21) for seismic waves, we see that the solution for a component E_i of the electric field is

$$E_i = E_0 \sin 2\pi\left(\frac{z}{\lambda} - ft\right), \tag{10.57}$$

where λ is the wavelength, f the frequency and $f\lambda = V$, the velocity of the wave.

The above considerations suggest that high-frequency electromagnetic waves travel in the ground in an analogous manner to seismic waves. Instead of being determined by the elastic parameters, the propagation of radar signals is dependent on the dielectric properties of the ground. A comparatively young branch of geophysical exploration has been developed to investigate underground structures with ground penetrating radar (GPR, or *georadar*). GPR makes use of the familiar "echo-principle" used in reflection seismology. A very short radar pulse, lasting only several nanoseconds (i.e., $\sim 10^{-8}$ s) is emitted by a mobile antenna on the ground surface. The path of the radar signal through the ground can be traced as a ray, which experiences refractions, reflections and diffractions at boundaries where the dielectric constant changes. A second antenna, the receiver, is located close to the transmitter, as in the case of seismic reflection, so as to receive near-vertical reflections from underground discontinuities. The signal-processing techniques of reflection seismology can also be applied to the georadar signal to help minimize the effects of diffractions and other noise. Consequently, georadar provides a detailed picture of the shallow subsurface structure (Fig. 10.28). It has become an important tool in environmental studies of near-surface features, such as buried and forgotten waste deposits, fracture patterns in otherwise uniform rock bodies, or the investigation of groundwater resources.

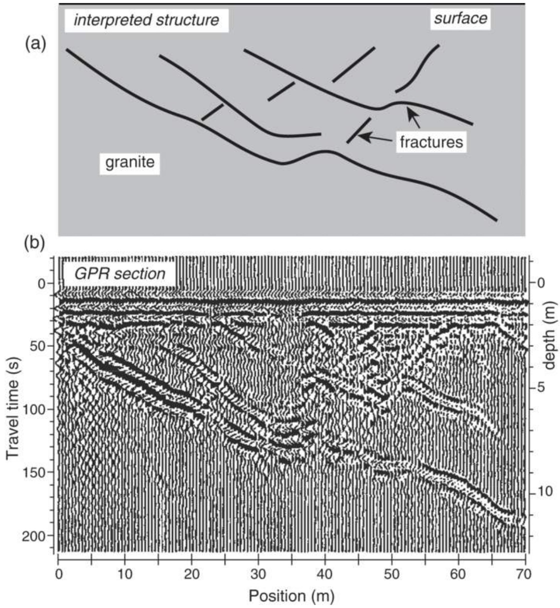


Fig. 10.28 Fracture pattern in a granitic bedrock revealed by ground penetrating radar: (a) geological cross-section; (b) processed georadar reflection section (courtesy of P. Annan).

High-frequency signals are rapidly attenuated with depth. On the one hand, geometrical spreading of the signal outward from its source (spherical divergence) causes a decrease of intensity with distance. More important is absorption of the signal by ground materials, which is a function of their conductivity. Depending on the composition of the soil or rocks (e.g., the presence of clay-rich layers or groundwater), the nature of subsurface structures and the frequency of the radar signal, the effective penetration may be up to 10 m, although conditions commonly restrict it to only a few meters. However, at a radar frequency of 10^8–10^9 Hz and with a velocity of $\sim 10^8$ m s_{-1} the resolution is in the range 0.1–1 m. Thus, despite its limited depth penetration, the high resolution of georadar makes it a powerful tool for near-surface geophysical exploration.

10.7 Electrical Conductivity in the Earth

The complicated structure of the crust and upper mantle results in large lateral variations in electrical conductivity. Apart from the oceans, sediments, and individual anomalous conductors, the outer carapace of the Earth is generally a poor electrical conductor. The physical mechanism of conductivity in silicate rocks is by semi-conduction, which can take place in three different ways (Section 10.2.3). Each type of semi-conduction is governed by a thermally activated process, in which the conductivity σ at temperature T is given by

$$\sigma = \sigma_0 e^{-E_a/kT}, \tag{10.58}$$

where k is Boltzmann's constant ($k = 1.38065 \times 10^{-23}$ J K^{-1}). The constant σ_0 is the hypothetical maximum value of the conductivity, reached asymptotically at very high temperatures. E_a is the activation energy of the particular type of semi-conduction. Its value determines the temperature range in which the thermally activated process becomes effective as a mechanism for σ. In the crust and upper mantle (i.e., in the lithosphere) impurity semi-conduction is likely the main mechanism in dry rocks. Electronic semi-conduction is probably dominant in the asthenosphere and deeper regions of the mantle. Ionic semi-conduction is an important mechanism at high temperature, but is unlikely to be significant below about 400 km, because it is suppressed by the high pressure at greater depth.

The electrical conductivity in the Earth at great depths is inferred from four sources: deep electrical sounding, geomagnetic variations, secular variations and extrapolation from laboratory experiments. The first two methods are based on induction effects arising from changes in the external part of the geomagnetic field; these encompass a broad spectrum with peaks of energy at several periods (see Fig. 10.7). Electrical and magnetotelluric sounding use the components with periods from milliseconds to one or two days. The inversion of MT data gives a conductivity pattern that is generally concordant with seismic data and related to the broad geological structure of the crust and upper mantle.

The time spectrum of external geomagnetic field variations contains some prominent periods that are longer than a day (Fig. 10.7). The study of the longer-period geomagnetic variations provides information about conductivity in the Earth down to about 2000 km. The longer the period of the variation the deeper its penetration depth. The daily (or diurnal) variation (Section 11.2.3.3) yields conductivity information to about 900 km. Magnetic storms last several days or weeks and have a strong 48-hour component, which is used to extend conductivity information to about 1000 km. In addition to the spectrum of geomagnetic variation shown in Fig. 10.7 there is a longer-period component related to the 11-year sunspot cycle. It results from increased solar activity and is accompanied by solar flares (see Section 11.3.1.1) and emissions of charged particles that augment the solar wind and excite ionospheric activity. Analysis of the 11-year component allows the model of mantle conductivity to be extended to about 2000 km depth. Our knowledge of the electrical conductivity in the mantle at depths greater than 2000 km cannot be obtained from analysis of effects related to the external magnetic field.

The secular variation of the internal geomagnetic field (Section 11.2.5) originates in the upper part of the fluid outer core. It consists of fluctuations in intensity and direction with periods of the order of $10–10^4$ years. If the secular variation could be observed at the core–mantle boundary it would be possible to determine conductivity throughout the mantle. Unfortunately, secular variation must be observed at the Earth’s surface, after it has passed through the conducting mantle, which acts as a filter. The signal is attenuated by the skin effect, preferentially affecting the highest frequencies. Thus, observations of high-frequency changes in secular variation place an upper limit on the average conductivity of the mantle, because a greater conductivity would block them out. From time to time, abrupt changes in the rate of secular variation take place for unknown reasons. A conspicuous example of these “geomagnetic jerks” occurred in 1969–1970, when a pulse in secular variation occurred with an estimated duration of less than two years. Although not all analysts concur, the effect is widely believed to be of internal origin. Analysis of the propagation of a secular-variation pulse provides an estimate of the mean conductivity of the whole mantle, which, integrated with data from other sources, gives the conductivity in the lower mantle.

Some of the different models of mantle conductivity that have been proposed are shown in Fig. 10.29. Differences between the models reflect increases in quantity and improvements in quality of geomagnetic data as well as advances in the techniques of data-processing, especially the development of inversion methods. Although the models diverge in many respects, they have some features in common. The conductivity averages about 10^{-2} S m^{-1} in the lithosphere and increases with increasing depth. Sharper rates of increase are found at depths of 410 and 660 km, where the olivine–spinel and spinel–perovskite phase changes occur, respectively (see Section 7.4.4.1). At about 700 km depth each model gives a conductivity of about 1 S m^{-1}, which rises in the lower mantle to be about 10–200 S m^{-1} at the core–mantle boundary. The secular variation is not uniform over the Earth’s surface. Large areas of continental size (the Central Pacific is the best studied) are characterized by slow rates of variation. This is possibly due to the additional screening effect of features in the D″ layer above the core–mantle boundary (Section 7.4.4.2), in which the conductivity may increase with depth to a value that is 100 times higher than in the overlying mantle.

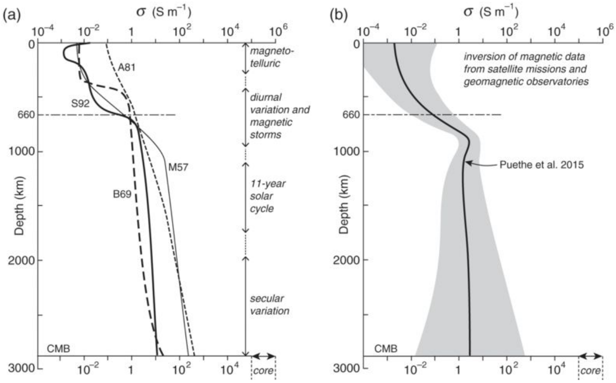


Fig. 10.29 (a) Models of electrical conductivity (σ) at depth in the mantle: M57, MacDonald, 1957; B69, Banks, 1969; A81, Achache et al., 1981; and S92, Stacey, 1992 (after Stacey, 1992). (b) Electrical conductivity profile of the mantle from inversion of data from the CHAMP and SWARM satellites and geomagnetic observatorie.

Source: Püthe, C., Kuvshinov, A., Khan, A. and Olsen, N. 2015. A new model of Earth's radial conductivity structure derived from over 10 yr of satellite and observatory magnetic data. *Geophys. J. Int.*, 203, 1864–1872.

Conductivity in the outer core is estimated by extrapolation from laboratory experiments and by first principle calculations. The core has the composition of an iron alloy, with an estimated iron content of ~80% and oxygen and silica as the alloying elements. Different compositional combinations indicate that the electrical conductivity in the outer core is around 1.1×10^{6} S m^{-1}.

10.8 *Suggestions for Further Reading*

Introductory Level

Kearey, P., Brooks, M., and Hill, I. 2002. *An Introduction to Geophysical Exploration* (3rd ed.). Oxford: Blackwell Publishing.

Intermediate and Advanced Level

Dobrin, M. B. and Savit, C. H. 1988. *Introduction to Geophysical Prospecting* (4th ed.). New York: McGraw-Hill.

Parasnis, D. S. 1997. *Principles of Applied Geophysics* (5th ed.). London: Chapman and Hall.

Reynolds, J. M. 2011. *An Introduction to Applied and Environmental Geophysics* (2nd ed.). Chichester: Wiley-Blackwell.

Sharma, P. V. 1997. *Environmental and Engineering Geophysics*. Cambridge: Cambridge University Press.

Telford, W. M., Geldart, L. P., and Sheriff, R. E. 1990. *Applied Geophysics*. Cambridge: Cambridge University Press.

10.9 Review Questions

1. Which characteristics of the ground determine its electrical resistivity?

2. Explain why geoelectrical resistivity measurements yield only an *apparent* resistivity.

3. What are telluric currents? How do they originate?

4. What is meant by the *skin depth* for the propagation of electromagnetic waves?

5. What are the *in-phase* and *quadrature* components in an electromagnetic induction survey? What causes the phase shift? Which component responds more strongly to the presence of a good conductor?

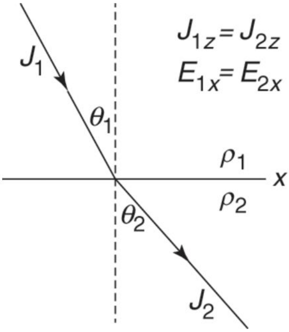
6. What is the magnetotelluric method of electromagnetic surveying? What are the merits of this method for deep Earth sounding?

7. What is ground penetrating radar and in which part of the electromagnetic spectrum is it operative? Why can ground penetrating radar signals be processed analogously to reflected and refracted seismic waves?

8. Why is ground penetrating radar a powerful method for exploring shallow subsurface structure? Which properties of the ground determine the effectiveness of the method and limit its depth range?

10.10 Exercises

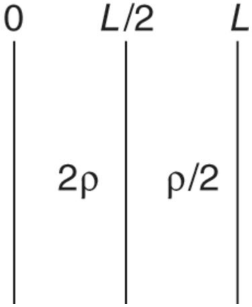
1. At the interface between two layers with electrical resistivities ρ_1 and ρ_2, as in the figure below, the electrical boundary conditions are: (1) the component of current density J_z *normal* to the interface is continuous, and (2) the component of electric field E_x *tangential* to the interface is continuous. A current flow-line makes angles θ_1 and θ_2 before and after refraction, respectively.



Derive the electrical “law of refraction” given by Eq. (10.32):

$$\frac{\tan\theta_1}{\tan\theta_2} = \frac{\rho_2}{\rho_1}.$$

2. What is the effective resistivity of a slab of thickness L composed of two half-slabs each of thickness $L/2$ and with resistivities (2ρ) and ($\rho/2$), respectively, as in the diagram?



3. Seawater is contaminating an aquifer that is the source of drinking water for a seaside town. The following measurements of apparent resistivity (ρ_a) were made at various electrode separations (a) with the expanding-spread Wenner method to investigate the leak.

a (m)	ρ_a (Ω m)	a (m)	ρ_a (Ω m)	a (m)	ρ_a (Ω m)
10	29.0	140	19.8	280	8.7
20	28.9	160	18.0	300	7.8
40	28.5	180	16.3	320	7.1
60	27.1	200	14.5	340	6.7
80	25.3	220	12.9	360	6.5
100	23.5	240	11.3	400	6.4
120	21.7	260	9.9	440	6.4

(a) Estimate the electrical resistivity of each layer.

(b) Divide the apparent resistivity at each position by the resistivity of the upper layer, then plot the normalized resistivity against electrode separation on a log–log diagram on the same scale as the model curves in Fig. 10.15.

(c) Match the measured curve with the model curves and estimate the depth to the interface.

4. In the Schlumberger resistivity method the separation of the current electrodes L is much larger than the separation a of the voltage electrodes. Suppose that the mid-point of the voltage pair is displaced by a distance x from the mid-point of the current electrode pair. Show that, for $(L - 2x) \gg a$, the apparent resistivity is given by

$$\rho_a = \frac{\pi}{4}\frac{V}{I}\frac{(L^2 - 4x^2)^2}{a(L^2 + 4x^2)}.$$

5. In the double-dipole resistivity method it is common to keep the separation of the pairs L an integer multiple n of the distance *a* between the electrodes in each pair, i.e., $L = na$.

(a) Rewrite the formula for the apparent resistivity with this assumption.

(b) If L is very large compared to a, modify the formula to show that the apparent resistivity is proportional to n^3.

6. Consider a double-dipole configuration in which the electrode pairs are not collinear but are broadside to each other (i.e., normal to the line joining them). The electrode separation is a and the distance between the mid-points of the pairs is $L = na$. Show that, for large values of n, the apparent resistivity is given in this case by

$$\rho_a = 2\pi n^3 a \frac{V}{I}.$$

7. Calculate the velocity of an electromagnetic wave in (a) basalt (dielectric constant $\kappa = 12$) and (b) water ($\kappa = 80.4$).

8. Calculate the "skin depths" of penetration in (a) granite ($\rho = 5000\ \Omega$ m) and (b) a pyrrhotite orebody ($\rho = 5 \times 10^{-5}\ \Omega$ m) for electromagnetic waves in surveys employing (1) electromagnetic induction ($f = 1$ kHz) and (2) ground penetrating radar ($f = 100$ MHz). Would these methods detect the conducting bodies if they were buried under a water-saturated soil layer, 3 m thick with resistivity 100 Ω m?

10.11 Computational Exercises

Jupyter notebook computational exercises can be found online for this chapter at www.cambridge.org/FoG3ed (Four-electrode methods).

11

The Earth's Magnetic Field

Preview

The geomagnetic field is generated by complex motions of electrically conducting liquid iron in the Earth's outer core. It has been studied for centuries at observatories and for decades from orbiting satellites. Outside the Earth, the magnetic field deflects the solar wind, a stream of charged particles from the Sun, thereby shielding the planet from harmful radiation. Some radiation penetrates deep into the outer atmosphere, where it ionizes air molecules. These cause the aurora and also form spherical shells of ions encircling the Earth that produce their own magnetic fields. We have learned about the magnetic fields of the other planets from earth-bound observations and from space missions. The geomagnetic field provides an important tool for exploring the mineral wealth of our planet. We learn how to interpret the shapes of the magnetic anomalies of simple geometric bodies, including block models of oceanic magnetic anomalies.

11.1 Magnetism

11.1.1 The Discovery of Magnetism

The earliest observations of magnetism were made before accurate records of discoveries were kept, so it is impossible to be sure of historical precedents. Around 800 BC, Greek philosophers wrote about the curious attractive properties of lodestone, a naturally occurring form of the mineral magnetite. In the fifth century BC, the Greek state of Thessaly founded a colony in what is now the Turkish province of western Anatolia called *Magnesia*. Nearby the Greeks found a ready supply of lodestone, pieces of which subsequently became known by the Latin word *magneta*; from this the term magnetism derives. The Greek philosophers ascribed the attractive forces of lodestone to metaphysical powers, which gave rise to false conceptions that persisted for centuries. Foremost among these was the view that magnetic forces were related to invisible fluids. This view persisted well into the nineteenth century. The power of a magnet seemed to flow from one pole to the other along lines of force that could be made visible by sprinkling a powder of iron filings on a paper held over the magnet. The term “flux” (synonymous with flow) is still found in “magnetic flux density,” which is regularly used as an alternative to “magnetic induction” for the fundamental magnetic field vector **B**.

Early in the Han dynasty, between 300 and 200 BC, Chinese artisans fashioned a rudimentary compass out of lodestone. It consisted of a spoon-shaped object, whose bowl balanced and could rotate on a flat polished surface. This compass may have been used in the search for gems and in the selection of sites for houses. By AD 1000 the Chinese had developed suspended and pivoted needle compasses. Their directional property led to the use of compasses for navigation long before the origin of the aligning forces was understood. As late as the twelfth century, the alignment of a compass was ascribed to its attempt to follow the pole star. It was later shown that the compass alignment was produced by a property of the Earth itself. Subsequently, the characteristics of terrestrial magnetism played an important role in advancing the understanding of magnetism.

11.1.2 Pioneering Studies in Terrestrial Magnetism

In 1269 a medieval scholar with the Latin nom-de-plume of Petrus Peregrinus, wrote the earliest known treatise of experimental physics (*Epistola de Magnete*). In it he summarized magnetic behavior known to that time and described simple laws of magnetic attraction. He introduced the concept of magnetic poles based on an experiment in which he placed a spherical magnet made of lodestone on a flat slab of iron and traced the lines of direction that the magnet assumed. These lines circled the lodestone sphere like geographical meridians and converged at two antipodal points, which Peregrinus called the *poles* of the magnet, by analogy to the geographical poles.

By AD 500 Chinese scholars were aware that magnetic compasses did not point exactly to geographical north, as defined by the stars. The angle between magnetic north and true north is called the magnetic *declination*. By the fourteenth century, the ships of the British navy were equipped with a mariner's compass, which became an essential tool for navigation. It was used in conjunction with celestial methods, and gradually it became apparent that the declination changed with position on the globe. The accumulation of mariners' observations during the fifteenth and sixteenth centuries established the worldwide pattern of declination.

In 1600 William Gilbert, an English scientist and physician to Queen Elizabeth, published *De Magnete*, a landmark treatise in which he summarized all that was then known about magnetism, including the results of about 17 years of his own research. Using small magnetic needles placed on the surface of a sphere of lodestone to study its magnetic field, he recognized the poles, where the needles stood on end, and the equator, where they lay parallel to the surface. Gilbert achieved the leap of imagination that was necessary to see the analogy between the attraction of the lodestone sphere and the known magnetic properties of the Earth. He recognized that the Earth itself behaved like a large magnet. This was the first unequivocal recognition of a geophysical property, preceding the laws of gravitation in Newton's *Principia* by almost a century. Although founded largely on qualitative, rather than quantitative, observations, *De Magnete* was the most important work on magnetism until the nineteenth century.

The discovery that the declination of the geomagnetic field changed with time was made by Henry Gellibrand, an English mathematician and astronomer, in 1634. He noted, on the basis of just three measurements made by William Borough in 1580, Edmund Gunter in 1622, and himself in 1634, that the declination had decreased by about 7° in this time. From these few observations he inferred what is now called the *secular variation* of the field.

Gradually the variation of the terrestrial magnetic field over the surface of the Earth was established. From 1698 to 1700 Edmund Halley, an English astronomer and mathematician, carried out an oceanographic survey with the prime purpose of studying compass variations in the Atlantic Ocean. This resulted in the publication of the first declination chart in 1702.

In 1785 Charles Augustin de Coulomb (1736–1806), a French scientist, published quantitative measurements of electrostatic and magnetic properties made with a torsion-balance he had invented. He established the inverse-square law of attraction and repulsion between small electrically charged balls. Using thin, magnetized steel needles about 24 inches (61 cm) in length, he also established that the attraction or repulsion between their poles varied as the inverse square of their separation. In the 1820s experiments conducted independently by Hans Christian Ørsted (1777–1851) and André Marie Ampère (1775–1836) – Danish and French scientists, respectively – established that an electrical current produces a magnetic force. The English experimenter Michael Faraday (1791–1867) showed in 1831 that the change of magnetic flux in a coil induced an electrical current in the coil.

The strengths of magnetic forces were related to mechanical units by the German scientist and mathematician Carl Friedrich Gauss (1777–1855), who assumed that static magnetism was carried by magnetic “charges” (now called “poles”), analogous to the carriers of static electricity. Gauss was responsible for setting up geomagnetic observatories, so that by 1837 global charts of the total intensity, inclination and declination were in existence. To analyze the dataset, Gauss applied the mathematical technique of spherical harmonic analysis, and thereby established that the main part of the Earth’s magnetic field was a dipole field that originated inside the Earth.

Since man first became aware of magnetic behavior, students of terrestrial magnetism have made important contributions to the understanding of magnetism as a physical phenomenon. In turn, advances in the physics of magnetism have helped geophysicists to understand the morphology and origin of the Earth’s magnetic field, and to apply this knowledge to geological processes, such as global tectonics. The physical basis of magnetism is fundamental to the geophysical topics of geomagnetism, rock magnetism, and paleomagnetism.

11.1.3 The Magnetic Dipole

Early investigators conceptualized gravitational, electrical, and magnetic forces between objects as instantaneous effects that took place through direct action-at-a-distance. Faraday introduced the concept of the *field* of a force as a property of the space in which the force acts. The force-field plays an intermediary role in the interaction between objects. For example, an electric charge is surrounded by an electrical field that acts to produce a force on a second charge. The pattern of a field is portrayed by *field lines.* At any point in a field the direction of the force is tangential to the field line and the intensity of the force to the number of field lines per unit cross-sectional area.

Historically, the laws of magnetism were established by relating the magnetic field to fictitious centers of magnetic force called magnetic poles, defined by comparison with the properties of a bar magnet. Gauss' work in 1832 showed that, in contrast to electrostatic charges, free magnetic poles cannot exist; each positive pole must be paired with a corresponding negative pole. The most important type of magnetic field – and also the dominant component of the geomagnetic field – is that of a *magnetic dipole*. This is the field of two magnetic poles of opposite sense that are infinitesimally close to each other. The geometry of the field lines shows the paths along which a free magnetic pole would move in the vicinity of the dipole (Fig. 11.1a). A tiny current loop (Fig. 11.1b) and a uniformly magnetized sphere (Fig. 11.1c) also have dipole-type magnetic fields around them. Although magnetic poles do not exist physically, many problems that arise in geophysical situations can be understood by assuming surface distributions of *fictitious* poles or dipoles.

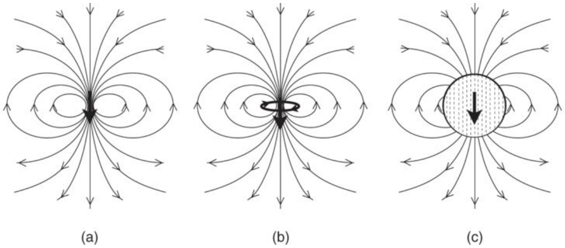


Fig. 11.1 The characteristic field lines of a magnetic dipole are found around (a) a short bar magnet, (b) a small loop carrying an electric current, and (c) a uniformly magnetized sphere.

Box 11.1 Potential of a Magnetic Dipole

Coulomb showed in 1785 that the force between the ends of two long, thin magnets is inversely proportional to the square of their separation. By analogy to electrical laws, Gauss in 1832 attributed the forces of magnetic attraction and repulsion to fictitious magnetic "charges," or poles. These observations resulted in an inverse square law for the force F between magnetic poles with strengths p_1 and p_2 at distance r from each other:

$$F(r) = K\frac{p_1 p_2}{r^2}, \tag{1}$$

where K is a proportionality constant.

Although magnetic poles are fictitious, many magnetic properties can be described and magnetic problems solved by assuming their existence. The magnetic field B of a magnetic pole p is defined as the force it exerts on a unit pole at distance r. From Eq. (1)

$$B(r) = K\frac{p}{r}. \quad (2)$$

The proportionality constant K is not dimensionless but is found to have the value $\mu_0/4\pi$, where μ_0 is an important physical constant, the *permeability* of free space. Also called the *magnetic constant,* it is one of the fundamental constants of physics, defined to be exactly equal to $4\pi \times 10^{-7}$ N A^{-2} (or henry/meter, H m^{-1}).

The magnetic potential W at a distance r from a pole of strength p is obtained by computing the work that would have to be expended *against* the field to move the unit mass from r to infinity. Using the magnetic field of a single pole given by Eq. (2) and the value $\mu_0/4\pi$ for K, the magnetic potential at distance r is

$$W = -\int_r^\infty B\mathrm{d}r = \frac{\mu_0 p}{4\pi r}. \quad (3)$$

Let a positive magnetic pole (or charge) be located at distance d from a negative pole, as in Fig. B11.1.1. At the point with polar coordinates (r, θ) the distances from the respective poles are r_+ and r_- and the magnetic potential at that point of the pair of opposite poles is the sum of the potentials of the positive and negative poles

$$W = \frac{\mu_0 p}{4\pi}\left(\frac{1}{r_+} - \frac{1}{r_-}\right),$$

(4)

$$W = \frac{\mu_0 p}{4\pi}\left(\frac{r_- - r_+}{r_+ r_-}\right).$$

(5)

In a *dipole* the separation of the poles is infinitesimally small compared to the distance to the point of observation (i.e., $d \ll r$). In this case, we get the approximate relations

$$r_+ \approx r - \frac{d}{2}\cos\theta$$

$$r_- \approx r + \frac{d}{2}\cos\theta'.$$

(6)

When $d \ll r$, we can write $\theta \approx \theta'$ and terms of order $(d/r)^2$ and higher can be neglected. This leads to the further simplifications

$$(r_- - r_+) \approx \frac{d}{2}(\cos\theta' + \cos\theta) \approx d\cos\theta$$

$$(r_+ r_-) \approx r^2 - \frac{d^2}{4}\cos^2\theta \approx r^2.$$

(7)

Substituting these values into Eq. (5) gives the dipole potential at the point (r, θ):

$$W = \frac{\mu_0}{4\pi} \frac{(pd)\cos\theta}{r^2}, \tag{8}$$

$$W = \frac{\mu_0}{4\pi} \frac{m\cos\theta}{r^2}. \tag{9}$$

The quantity $m = (pd)$ is known as the *magnetic moment* of the magnet.

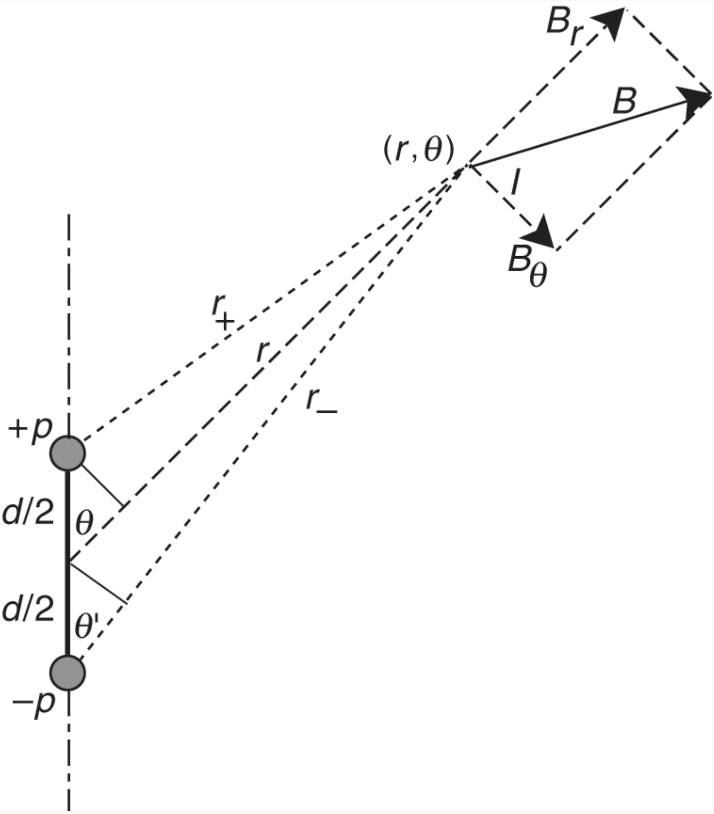


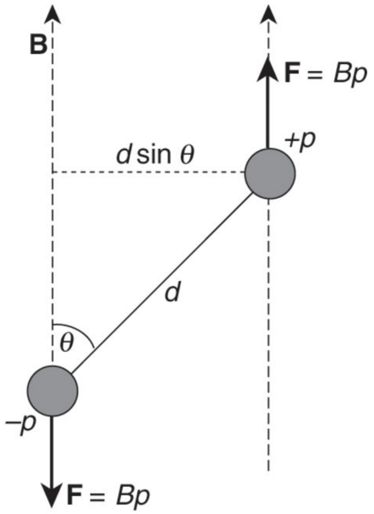
Fig. B11.1.1 Geometry for the calculation of the potential of a pair of magnetic poles.

The magnetic potential of a pair of equal and opposite poles, $+p$ and $-p$, located a distance d apart is derived in Box 11.1. The line joining the poles defines an axis, about which the field has rotational symmetry. The pair of opposite poles form a *dipole* when their separation becomes infinitesimally small compared to the distance to the point of observation (i.e., $d \ll r$). The potential of a magnetic dipole at the point (r, θ) is

$$W = \frac{\mu_0}{4\pi}\frac{m\cos\theta}{r^2}. \qquad (11.1)$$

The *magnetic moment* (m) of the dipole is defined from the observation that a bar magnet experiences a torque in a uniform magnetic field that turns the magnet parallel to the field direction. Consider the forces exerted by a uniform magnetic field B on a pair of magnetic poles of strength p separated by a distance d as in Fig. 11.2. A force equal to (Bp) acts on the positive pole and an equal and opposite force acts on the negative pole. If the magnetic axis is oriented at angle θ to the field, the perpendicular distance between the lines of action of the forces is $d \sin \theta$. The magnitude T of the torque on the magnet is

$$T = (pd)B \sin\theta = mB \sin\theta. \tag{11.2}$$



torque = $Fd\sin\theta$

$\mathbf{T} = pd\, B\sin\theta = \mathbf{m} \times \mathbf{B}$

Fig. 11.2 Definition of the magnetic moment **m** of a pair of magnetic poles.

The size of the torque is proportional to m, even when the separation of the poles becomes very small, as in the case of the dipole. The moment of the magnet, **m**, is a vector with direction along the axis of the dipole. Taking into account the direction of the torque and using the conventional notation for the cross product of two vectors, we obtain a vector equation for the torque **T** on a magnetic moment **m** in a field **B**:

$$\mathbf{T} = \mathbf{m} \times \mathbf{B}. \tag{11.3}$$

The same expression is obtained for the torque on a small loop (or coil) with area *A*, carrying an electrical current *I* in a magnetic field **B** (Box 11.2). In this case the vector **m** has magnitude (*IA*) and direction parallel to the normal to the plane of the current loop. It is evident that **m** – the strength of the dipole – corresponds to the magnetic moment of the current loop. At distances greater than the dimensions of the loop, the magnetic field is that of a dipole at the center of the loop (Fig. 11.1b). The definition of **m** in terms of a current-carrying loop shows that magnetic moment has the units of current times area, or ampère meter^2 (A m^2).

Box 11.2 Electrical Currents and Magnetic Fields

The equation used to define the magnetic **B**-field was formulated by Lorentz in 1879. Let *q* be an electrical charge that moves with velocity *v* through a magnetic field **B** (Fig. B11.2.1). The charged particle experiences a deflecting force **F** given by Lorentz's law, which in SI units is

$$\mathbf{F} = q(\mathbf{v} \times \mathbf{B}). \tag{1}$$

This equation defines the unit of the magnetic **B**-field. It is called a *tesla* and has the dimensions N A^{-1} m^{-1}. The electric field **E** is defined as the force acting on a unit charge, so we can rewrite the equation as

$$\mathbf{E} = \mathbf{v} \times \mathbf{B}. \tag{2}$$

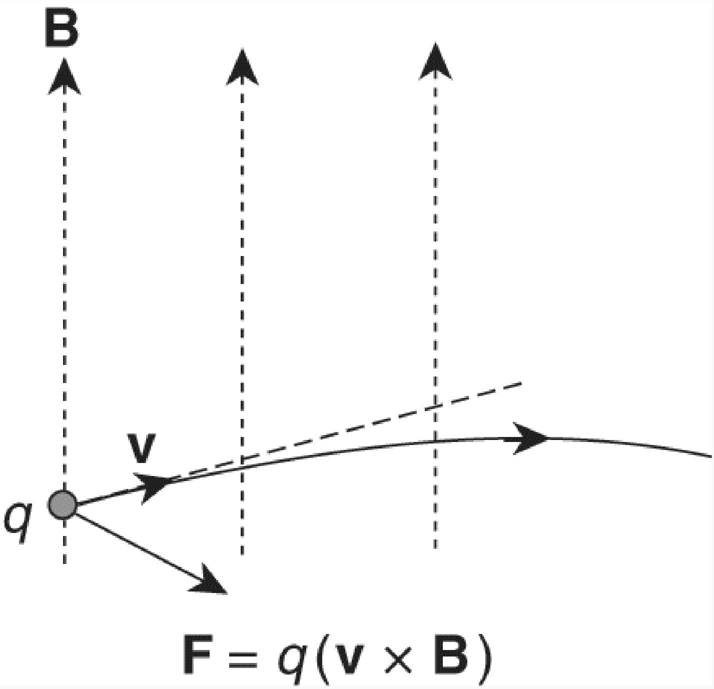


Fig. B11.2.1 Lorentz's law for the deflecting force **F** experienced by an electrical charge that moves with velocity **v** through a magnetic field **B**.

Imagine the moving charge to be confined to move along a conductor of length **dl** and cross-section *A* (Fig. B11.2.2). Let the number of charges per unit volume be *N*. The number inside the element **dl** is then *NA* **dl**. Each charge experiences a deflecting force given by Eq. (1). The velocity of each charge is in the direction **dl**, so the force acting on the element **dl** is

$$\mathrm{d}\mathbf{F} = (NA\mathrm{d}l)q(\mathbf{v} \times \mathbf{B}) = NAqv(\mathbf{dl} \times \mathbf{B}). \tag{3}$$

The electrical current *I* along the conductor is the total charge that crosses *A* per second, and is given by *I* = *NAqv*. From Eq. (3) we get the force acting on the element **dl** of a conductor carrying a current *I* in a magnetic field **B**:

$$\mathrm{d}\mathbf{F} = I(\mathrm{d}\mathbf{l} \times \mathbf{B}). \tag{4}$$

Equation (4) is the *Biot-Savart* law. It can be used to determine the torque exerted on a small rectangular loop PQRS in a magnetic field (Fig. B11.2.3a).

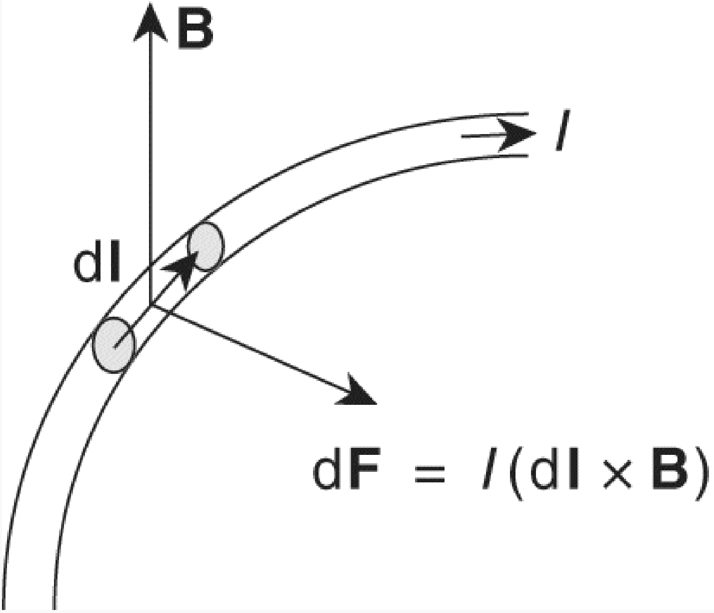


Fig. B11.2.2 Biot-Savart law for the force experienced by an element d**l** of a conductor carrying a current I in a magnetic field **B**.

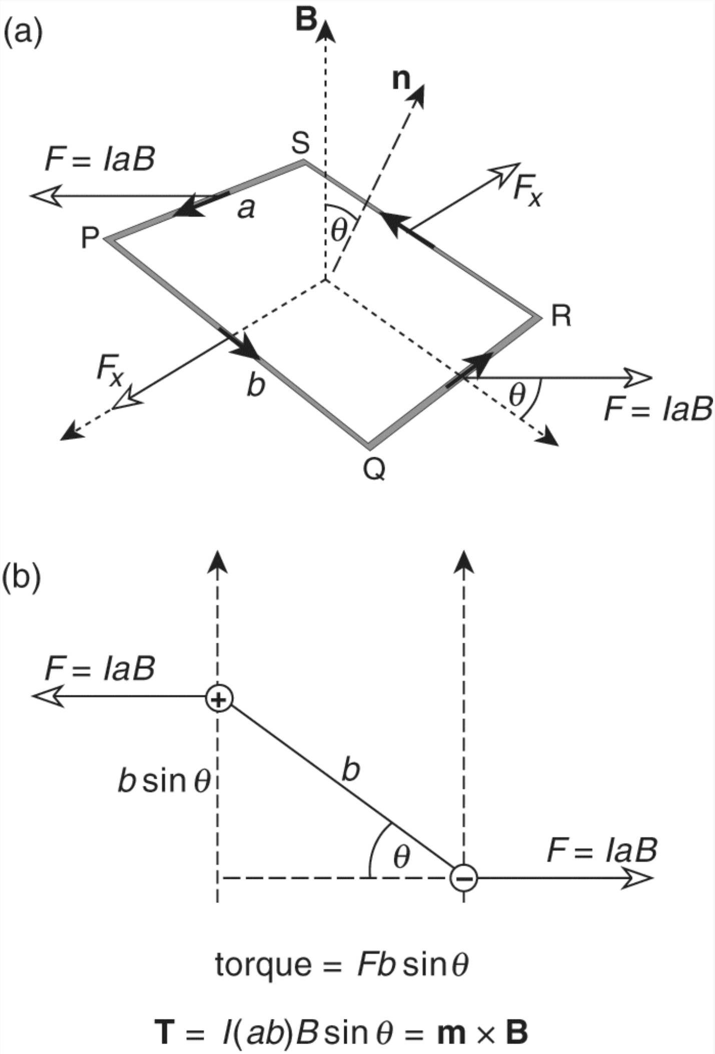

(a)
B
n
S
F = IaB
a
P
θ
F_x
R
F_x
b
θ
F = IaB
Q
(b)
F = IaB
+
b sin θ
b
θ
−
F = IaB
torque = Fb sinθ
T = I(ab)B sin θ = m × B

Fig. B11.2.3 (a) Forces on a rectangular loop carrying a current *I* in a uniform magnetic field *B*. (b) Derivation of the torque τ experienced by the loop.

Let the lengths of the sides of the loop be a and b, respectively, and define the x-axis parallel to the sides of length a. The area of the loop can be expressed as a vector with magnitude $A = ab$, and direction **n** normal to the plane of the loop. Suppose that a current I flows in the loop and that a magnetic field **B** acts normal to the x-axis, making an angle θ with the normal to the plane of the loop. Applying Eq. (3), a force F_x equal to ($IbB \cos \theta$) acts on the side PQ in the direction of $+x$; its effect is canceled by an equal and opposite force F_x acting on side RS in the direction of $-x$. Forces equal to (IaB) act in opposite directions on the sides QR and SP (Fig. B11.2.3b). The perpendicular distance between their lines of action is $b \sin \theta$, so the torque T experienced by the current loop is

$$T = (IaB)b \sin\theta = (IA)B \sin\theta. \tag{5}$$

This yields the vector equation

$$\mathbf{T} = \mathbf{m} \times \mathbf{B}. \tag{6}$$

The vector **m** has magnitude (IA) and direction parallel to **n**, the normal to the plane of the current loop. This expression is valid for an arbitrary small current loop of area A, regardless of its shape.

11.2 Geomagnetism

11.2.1 The Geomagnetic Elements

The intensity of the Earth’s magnetic field is measured in *tesla* (Box 11.2). However, a tesla is an extremely strong magnetic field, such as we might observe between the poles of an electromagnet. The Earth’s magnetic field is much weaker. Its maximum intensity occurs near to the magnetic poles, where it amounts to about 6×10^{-5} T. Modern instruments for measuring magnetic fields (called magnetometers) have a sensitivity of about 10^{-9} T; this unit is called a nanotesla (nT) and has been adopted in geophysics as the practical unit for expressing the geomagnetic field intensity. Most geomagnetic surveys carried out until the 1970s used the now superseded c.g.s. system of units, in which the magnetic field was measured in *gauss* (G), a unit equivalent to 10^{-4} T. The practical unit of magnetic exploration was 10^{-5} gauss, called a *gamma* (γ). The former unit (γ) is therefore equal to 10^{-9} T, which is the new unit (nT).

The magnetic vector can be expressed in Cartesian components parallel to any three orthogonal axes. The *geomagnetic elements* are field components parallel to the geographic north and east directions and the vertically downward direction (Fig. 11.3). Alternatively, the geomagnetic elements can be expressed in spherical polar coordinates. The magnitude of the magnetic vector is given by the field-strength F; its direction is specified by two angles. The *declination* D is the angle between the magnetic meridian and the geographic meridian; the *inclination* I is the angle at which the magnetic vector dips below the horizontal. The Cartesian (X, Y, Z) and spherical polar (F, D, I) sets of geomagnetic elements are related to each other as follows:

$$\begin{aligned}
&X = F\cos I\cos D \quad Y = F\cos I\sin D \qquad Z = F\sin I\\
&F = \sqrt{X^2+Y^2+Z^2}\\
&D = \arctan(Y/X) \quad I = \arctan\left(Z/\sqrt{X^2+Y^2}\right).
\end{aligned} \tag{11.4}$$

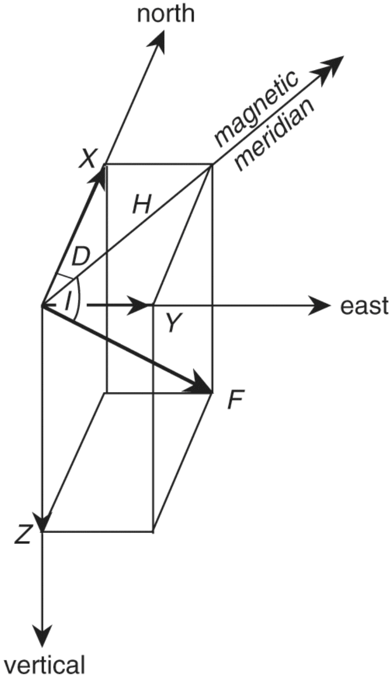

north
magnetic meridian
X
H
D
I
east
Y
F
Z
vertical

Fig. 11.3 Definition of the geomagnetic elements. The geomagnetic field can be described by north (*X*), east (*Y*) and vertically downward (*Z*) Cartesian components, or by the angles of declination (*D*) and inclination (*I*) together with the total field intensity (*F*).

11.2.2 Separation of the Magnetic Fields of External and Internal Origin

The magnetic field and its potential W at any point can be expressed in terms of the spherical polar coordinates (r, θ, ϕ) of the point of observation. Gauss expressed the potential of the geomagnetic field as an infinite series of terms in these coordinates. Essentially, his method divides the field into separate components that decrease at different rates with increasing distance from the center of the Earth. The detailed analysis is complicated and beyond the scope of this text. Analogous to the treatment of the gravitational potential (see Box 3.4), the magnetic potential W can be expressed as

$$W = R\sum_{n=1}^{\infty}\left(A_n r^n + \frac{B_n}{r^{n+1}}\right)\sum_{m=0}^{n} Y_n^m(\theta, \phi), \quad (11.5)$$

where R is the Earth's radius.

This is a rather formidable expression, but fortunately the most useful terms are quite simple. The summation signs indicate that the total potential is made up of an infinite number of terms with different values of n and m. The expression in parentheses describes the variation of the potential with distance r. For each value of n there will be different dependencies (e.g., on r, r^2, r^3, r^{-2}, r^{-3}, etc.). The function $Y_n^m(\theta, \phi)$ describes the variation of the potential when r is constant, i.e., on the surface of a sphere. It is called a *spherical harmonic function*, because it has the same value when θ or ϕ is increased by an integral multiple of 2π. For observations made on the spherical surface of the Earth, the constants A_n describe parts of the potential that arise from magnetic field sources outside the Earth, which are called the *geomagnetic field of external origin*. The constants B_n describe contributions to the magnetic potential from sources inside the Earth. This part is called the *geomagnetic field of internal* *origin*.

The potential itself is not measured directly. The geomagnetic field elements (Fig. 11.3) are recorded at magnetic observatories. Ideally these should be distributed uniformly over the Earth's surface, but in fact they are predominantly in the northern hemisphere. The geomagnetic field components are directional derivatives of the magnetic potential and depend on the same coefficients A_n and B_n. Observations of magnetic field elements at a large number of measurement stations with a worldwide distribution allows the relative importance of A_n and B_n to be assessed. From the sparse dataset available in 1838, Gauss was able to show that the coefficients A_n are very much smaller than B_n. He concluded that the field of external origin was insignificant and that the field of internal origin was predominantly that of a dipole.

11.2.3 The Magnetic Field of External Origin

The magnetic field of the Earth in space has been measured from satellites and spacecraft. The external field has a quite complicated appearance (Fig. 11.4). It is strongly affected by the *solar wind,* a stream of electrically charged particles (consisting mainly of electrons, protons, and helium nuclei) that is constantly emitted by the Sun. The solar wind is a *plasma.* This is the physical term for an ionized gas of low particle density made up of nearly equal concentrations of oppositely charged ions. At the distance of the Earth from the Sun (1 AU), the density of the solar wind is about 7 ions per cm^3, and it produces a magnetic field of about 6 nT. The solar wind interacts with the magnetic field of the Earth to form a region called the *magnetosphere.* At distances greater than a few Earth radii the interaction greatly alters the magnetic field from that of a simple dipole.

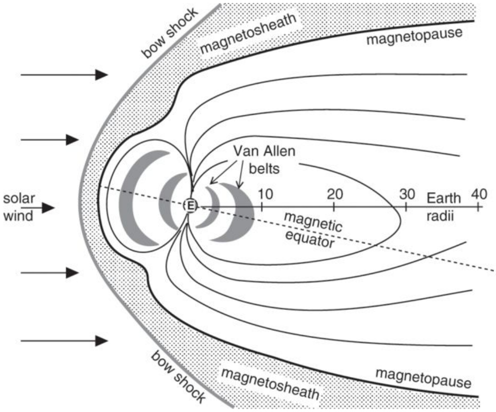

bow shock
magnetosheath
magnetopause
Van Allen
belts
solar
wind
E
10
20
30
Earth
radii
40
magnetic
equator
bow shock
magnetosheath
magnetopause

Fig. 11.4 Schematic cross-section through the magnetosphere, showing various regions of interaction between the Earth’s magnetic field and the solar wind.

The velocity of the solar wind relative to the Earth is about 450 km s^{-1}. At a great distance (about 15 Earth radii) from the Earth, on the day side, the supersonic solar wind collides with the thin upper atmosphere. This produces an effect similar to the build-up of a shock wave in front of a supersonic aircraft. The shock front is called the *bow shock* region (Fig. 11.4); it marks the outer boundary of the magnetosphere. Within the bow shock region the solar wind is slowed down and heated up. After passing through the shock front the solar wind is diverted around the Earth in a region of turbulent motion called the *magnetosheath*. The moving charged particles of the solar wind constitute electrical currents. They produce an *interplanetary magnetic field*, which reinforces and compresses the geomagnetic field on the day side and weakens and stretches it out on the night side of the Earth. This results in a geomagnetic tail, or *magnetotail*, which extends to great distances “downwind” from the Earth. The Moon’s distance from the Earth is about 60 Earth radii and so its monthly orbit about the Earth brings it in and out of the magnetotail on each circuit. The transition between the deformed magnetic field and the magnetosheath is called the *magnetopause*.

11.2.3.1 The Van Allen Radiation Belts

Charged particles that penetrate the magnetopause are trapped by the geomagnetic field lines and form the *Van Allen radiation belts*. These constitute two doughnut-shaped regions coaxial with the geomagnetic axis (Fig. 11.5). The inner belt contains mainly protons, the outer belt energetic electrons. Within each belt the charged particles move in helical fashion around the geomagnetic field lines (Fig. 11.6). The pitch of the spiraling motion gets smaller as the particle comes ever closer to the Earth and the field intensity increases; eventually it reaches zero and reverses sense. This compels the particles to shuttle rapidly from one polar region to the other along the field lines. The inner Van Allen belt starts about 1000 km above the Earth’s surface and extends to an altitude of about 3000 km (Fig. 11.5); the outer belt occupies a doughnut-shaped region at distances between about 3 and 4 Earth radii (20,000–30,000 km) from the center of the Earth.

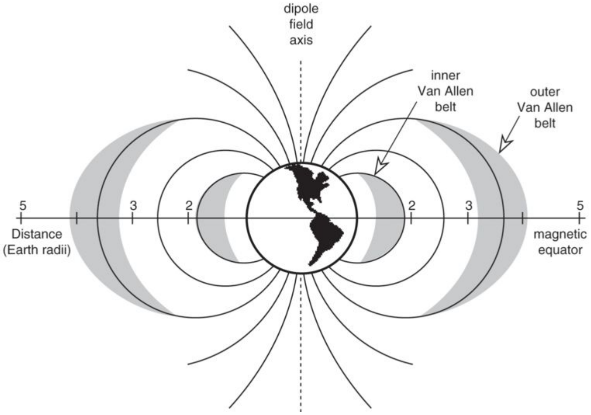

dipole
field
axis
inner
Van Allen
belt
outer
Van Allen
belt
5
3
2
2
3
5
Distance
(Earth radii)
magnetic
equator

Fig. 11.5 Schematic representation of the inner and outer Van Allen belts of charged particles trapped by the magnetic field of the Earth.

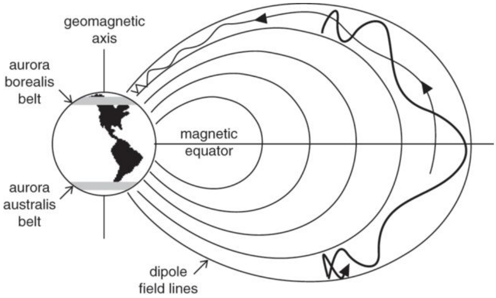

geomagnetic
axis
aurora
borealis
belt
aurora
australis
belt
magnetic
equator
dipole
field lines

Fig. 11.6 Charged particles from the solar wind are constrained to move in a helical fashion about the geomagnetic field lines.

11.2.3.2 The Ionosphere

The effects described above illustrate how the Earth's magnetic field acts as a shield against much of the extra-terrestrial radiation. The atmosphere acts as a protective blanket against the remainder. Most of the very short-wavelength fraction of the solar radiation that penetrates the atmosphere does not reach the Earth's surface. Energetic γ- and X-rays and ultraviolet radiation cause ionization of the molecules of nitrogen and oxygen in the thin upper atmosphere at altitudes between about 50 km and 1500 km, forming an ionized region called the *ionosphere*. It is formed of five layers, labeled the D, E, F1, F2, and G layers from the base to the top. Each layer can reflect radio waves. The thicknesses and ionizations of the layers change during the course of a day; all but one or two layers on the night side of the Earth disappear while they thicken and strengthen on the day side (Fig. 11.7). A radio transmitter on the day side can bounce signals off the ionosphere that then travel around the world by multiple reflections between the ground surface and the ionosphere. Consequently, radio reception of distant stations from far across the globe is best during the local night hours. The D layer is closest to the Earth at an altitude of about 80–100 km. It was first discovered in 1902, before the nature of the ionosphere was known, because of its ability to reflect long-wavelength radio waves, and is named the *Kennelly–Heaviside layer* in honor of its discoverers. The E layer is used by short-wave amateur radio enthusiasts. The F layers are the most intensely ionized.

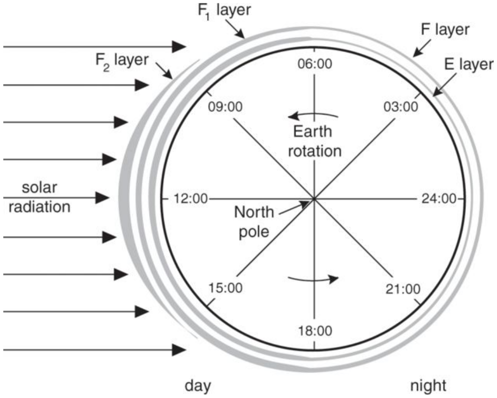


Fig. 11.7 Cross-section through the Earth showing the layered structure of the ionosphere.

Source: A.N. Strahler, *The Earth Sciences*, 681 pp., New York: Harper and Row, 1963.

11.2.3.3 Diurnal Variation and Magnetic Storms

The ionized molecules in the ionosphere release swarms of electrons that form powerful, horizontal, ring-like electrical currents. These act as sources of external magnetic fields that are detected at the surface of the Earth. The ionization is most intense on the day side of the

Earth, where extra layers develop. The Sun also causes atmospheric tides in the ionosphere, partly due to gravitational attraction but mainly because the side facing the Sun is heated up during the day. The motions of the charged particles through the Earth's magnetic field produce an electrical field, according to Lorentz's law (Box 11.2), which drives electrical currents in the ionosphere. In particular, the horizontal component of particle velocity interacts with the vertical component of the geomagnetic field to produce horizontal electrical current loops in the ionosphere. These currents cause a magnetic field at the Earth's surface. As the Earth rotates beneath the ionosphere the observed intensity of the geomagnetic field fluctuates with a range of amplitude of about 10–30 nT at the Earth's surface and a period of one day (Fig. 11.8a). This time-dependent change of geomagnetic field intensity is called the *diurnal* (or *daily*) *variation*.

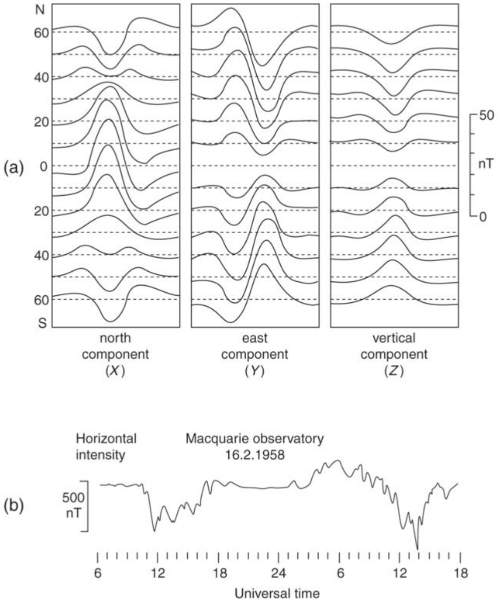


Fig. 11.8 (a) The time-dependent daily variation of the components of geomagnetic field intensity at different latitudes. Source: S. Chapman and J. Bartels, *Geomagnetism, Vol. I: Geomagnetic and Related Phenomena, Vol. II: Analysis of Data and Physical Theories*, 1049 pp., Oxford: Oxford University Press, 1940. (b) The variation of horizontal field intensity during a magnetic storm.

Source: T. Ondoh and H. Maeda, Geomagnetic storm correlation between the northern and southern hemisphere, *J. Geomag. Geoelectr*. 14, 22–32, 1962.

The magnitude of the diurnal variation depends on the latitude at which it is observed. Because it greatly exceeds the accuracy with which magnetic fields are measured during surveys, the diurnal variation must be compensated by correcting field measurements accordingly. The intensity of the effect depends on the degree of ionization of the ionosphere and is therefore determined by the state of solar activity, which is not constant. On days when the activity of the Sun is especially low, the diurnal variation is said to be of solar quiet (Sq) type. On normal days, or when the solar activity is high, the Sq variation is overlain by a solar disturbance (SD) variation. Solar activity changes periodically with the 11-year cycle of sunspots and solar flares. The enhanced emission of radiation associated with these solar phenomena increases the ionospheric currents. These give rise to rapidly varying, anomalously strong magnetic fields (called *magnetic storms*) with amplitudes of up to 1000 nT at the Earth's surface (Fig. 11.8b). The ionospheric disturbance also disrupts short-wave to long-wave radio transmissions. Magnetic surveying must be suspended temporarily while a magnetic storm is in progress, which can last for hours or days, depending on the duration of the solar activity.

11.2.4 The Magnetic Field of Internal Origin

To understand how geophysicists describe the geomagnetic field of internal origin mathematically, we return to Eq. (11.5). First, we follow Gauss and omit the coefficients A_n of the external field. The spherical harmonic functions $Y_n^m(\theta, \phi)$ that describe the variation of potential on a spherical surface are then written in expanded form (see Box 3.4). The potential *W* of the field of internal origin becomes

$$W = R\sum_{n=1}^{\infty}\sum_{m=0}^{n}\left(\frac{R}{r}\right)^{n+1}(g_n^m \cos m\phi + h_n^m \sin\theta)P_n^m(\cos\theta). \tag{11.6}$$

Here, *R* is the Earth’s radius, as before, and $P_n^m(\cos\theta)$ are called Schmidt polynomials, which are related to the associated Legendre polynomials.

Box 11.3 Multipole Representation of the Geomagnetic Field

The potential W of the magnetic field of internal origin is

$$W = R\sum_{n=0}^{n=\infty}\sum_{m=0}^{m=n}\left(\frac{R}{r}\right)^{n+1}(g_n^m \cos m\phi + h_n^m \sin m\phi)P_n^m(\cos\theta).$$

Each term in this expression is the potential of the magnetic field due to a particular combination of magnetic poles. For example, there are three terms for which $n = 1$, with corresponding coefficients g_1^0, g_1^1, and h_1^1. Each corresponds to a dipole field. The dipole g_1^0 is parallel to the axis of rotation of the Earth and the other two are at right angles to each other in the equatorial plane. The five terms with $n = 2$ describe a geometrically more complex *quadrupole* field. Just as a dipole field results when two single poles are brought infinitesimally close to each other (see Section 11.1.3), a quadrupole field results when two opposing coaxial dipoles are brought infinitesimally close end-to-end. As its name implies, the quadrupole field derives from four (2^2) magnetic poles of which two are "north" poles and two are "south" poles. The terms with $n = 3$ describe an *octupole* field which is characterized by eight (2^3) poles. The terms with $n = N$ describe a field that arises from a configuration of 2^N poles.

The configurations of axial dipole, quadrupole, and octupole fields relative to a reference sphere are illustrated in Fig. B11.3.1 for the case where the order m of the Gauss coefficients is zero. If $m = 0$, the potential is constant around a circle of "latitude" defined by a chosen combination of θ and r. This kind of field is said to have *zonal symmetry*. Any cross-section that contains the axis of symmetry is representative for the symmetry. Fig. B11.3.1 shows axial cross-sections of the simplest field line geometries corresponding to (a) dipole, (b) quadrupole, and (c) octupole fields; each field is rotationally symmetrical about the axis of the configuration. The corresponding *zonal spherical harmonics* are illustrated symbolically by shading the alternate zones in which magnetic field lines leave or return to the surface of a sphere.

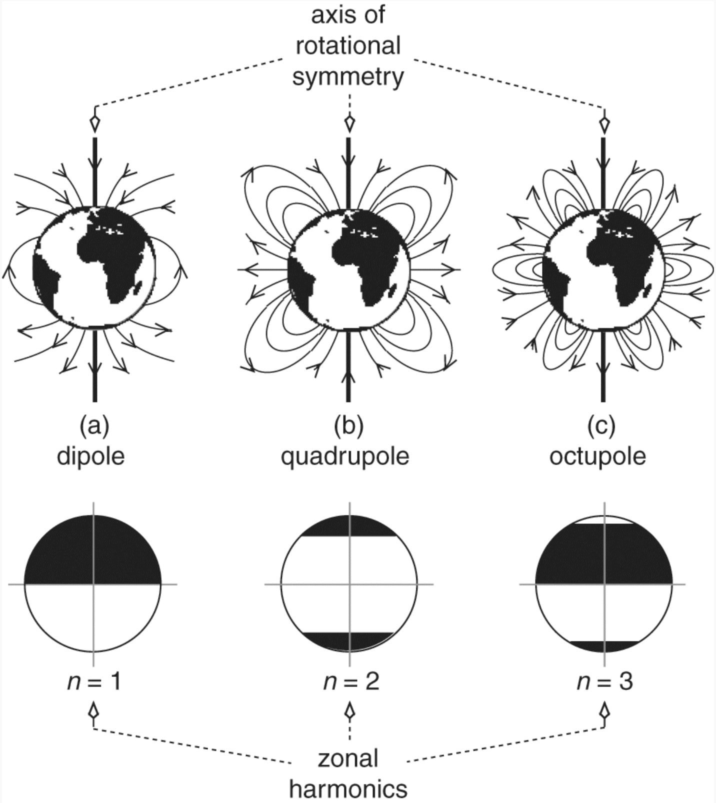


Fig. B11.3.1 Axial cross-sections showing the field line geometries of (a) dipole, (b) quadrupole, and (c) octupole fields; each field is rotationally symmetrical about the axis of the configuration. The corresponding zonal spherical harmonics are illustrated symbolically by shading the alternate zones in which magnetic field lines leave or return to the surface of a sphere. The lower part of the figure shows the symmetry relative to the axis of rotation of sectorial and tesseral spherical harmonic functions.

The dipole field is horizontal at the magnetic equator. In the southern hemisphere the field lines leave the reference sphere; in the north hemisphere they return to it. In the northern hemisphere the field of an axial quadrupole is horizontal at magnetic latitude 35.3 N. North of this latitude the field lines of the quadrupole leave the reference sphere. An equivalent circle of latitude is located in the southern hemisphere at 35.3 S; south of this latitude the quadrupole field lines also leave the Earth. The field lines re-enter the Earth in the band of latitudes between those where the field is horizontal. The symmetry of the quadrupole field is described by these three *zones* around the axis. The axial octupole field exhibits zonal symmetry with four zones; two zones in which the field leaves the Earth alternate with two zones in which it re-enters it.

The symmetry of *sectorial harmonics* for which the degree *n* and order *m* of the Gauss coefficients are equal (e.g., g_1^1 , h_1^1 , g_2^2 , h_2^2) is characterized by an even number of sectors around the equator in which the field lines alternately leave and re-enter the Earth (see also Box 3.4). The pattern of symmetry of the *tesseral harmonics* for which $m \neq n$ (e.g., g_2^1 , h_2^1 , g_3^2 , h_3^1) is defined by the intersections of circles of latitude and longitude, which bound alternating domains in which the field lines leave and re-enter the Earth.

Multipole representation allows the very complex geometry of the total field to be broken down into contributions from a number of fields with simple geometries. It must be kept in mind that multipole fields are not a physical expression of the magnetic field, because magnetic poles do not exist. It is a convenient technique for describing the field mathematically.

Equation (11.6) is a *multipole* expression of the geomagnetic potential. It relates the potential of the measured field to the potentials of particular combinations of magnetic poles (Box 11.3). The constants g_n^m and h_n^m in the geomagnetic potential are called the Gauss (or Gauss–Schmidt) coefficients of degree *n* and order *m*. Inspection of Eq. (11.6) shows that they have the same dimensions (nT) as the **B**-field. Their values are computed from analysis of measurements of the geomagnetic field.

Data from several sources are integrated in modern analyses of the Earth's magnetic field. Until the dawn of the satellite era, continuous records at magnetic observatories were the principal sources of geomagnetic data. Average values of the geomagnetic elements were determined, from which optimum values for the Gauss coefficients could be derived. Currently, about 200 permanent observatories make continuous measurements of the field. However, satellites orbiting the Earth in low near-polar orbits now provide most of the high-quality data used to model the field. The Polar Orbiting Geophysical Observatory (POGO), launched in 1965, was the first to deliver field measurements, but the greatest advance came with the Magnetic Field Satellite (MAGSAT), which delivered high-quality data during a six-month mission in 1979–1980. In 1999, a Danish satellite, ØRSTED, was placed in an elliptical, low polar orbit, with altitude between 650 km and 865 km. As on MAGSAT, it carried a vector magnetometer and a total field magnetometer, each with a sensitivity of 0.5 nT; the mission was dedicated to surveying the geomagnetic field. The German satellite, CHAMP (see also Section 3.5.4), was launched in 2000 in a lower, near-circular orbit on a planned five-year mission, which in fact lasted until 2010. In addition to measuring the gravity field, this satellite was also equipped to make scalar and vector measurements of the magnetic field. The low orbit and improved instrumentation were designed to provide magnetic measurements with high resolution.

The most recent satellite-borne magnetic survey, the SWARM Constellation Mission, was launched by the European Space Agency in 2013. Its successful initial four-year period of operation has been extended until at least 2021. SWARM consists of three identical satellites (*Alpha, Bravo*, and *Charlie*) flying together in near-polar orbits, each carrying a vector and a scalar magnetometer. Satellites *Alpha* and *Charlie* orbit the Earth side-by-side, separated horizontally by 1.4 degrees in longitude (155 km at the equator), at an initial altitude of 450 km; the third satellite, *Bravo*, orbits above this pair at

530 km altitude. The satellite orbits change with time during the mission, which allows the magnetometers to optimize sampling of the field. Data from the trio of satellites is managed in such a way that their outputs can be coupled in pairs, which enables them to be used as horizontal and vertical magnetic gradiometers (Section 11.4.3.2). This largely obviates the need to compensate for diurnal variations in the geomagnetic field and provides enhanced resolution of small-scale magnetic anomalies.

In principle, an infinite number of Gauss coefficients would be needed to define the field completely. A global model of the field is provided by the International Geomagnetic Reference Field (IGRF), which is updated every five years. The current version (IGRF-12, 2015) is based on Gauss coefficients up to degree and order 13. The IGRF also models a linear approximation to the rate of change of each of the Gauss coefficients (its *secular variation*), which permits correction of the current values between update years.

The Gauss coefficients become smaller with increasing degree n; this decrease provides a way of estimating the origin of the internal field. The analysis involves a technique called *power spectral analysis*. The distance across a feature of the magnetic field (for example, a region where the field is stronger than average) is called the wavelength of the feature. As in the case of gravity anomalies, deep-seated magnetic sources produce broad (long-wavelength) magnetic anomalies, while shallow sources result in narrow (short-wavelength) anomalies. Spectral analysis consists of calculating the power (alternatively called the energy density) associated with each “frequency” in the signal. This is obtained by computing the sum of the squares of all coefficients with the same degree. In the case of the geomagnetic field, the spectral analysis is based on the values of the Gauss coefficients. The spatial frequency of any part of the observed field is contained in the degree n of the coefficients. Low-degree terms (those with small values of n) correspond to long-wavelength features, high-degree terms are related to short-wavelength features.

Measurements of the geomagnetic field have been made from satellites that carried both a flux-gate vector magnetometer and a scalar magnetometer of the nuclear precession or optical pumping type (see Section 11.4.2.3). The MAGSAT mission in 1979–1980 measured the geomagnetic field from a nearly circular orbit at altitude 420 km with an accuracy high enough to enable a global spherical harmonic analysis that yielded Gauss coefficients up to degree $n = 66$. More recently, the CHAMP mission measured the geomagnetic field for a period of 10 years from 2000 to 2010. The satellite was placed in a nearly circular orbit at an initial altitude of 454 km above the Earth's surface, which progressively decreased to 300 km. The CHAMP data have been combined with marine magnetic and aeromagnetic data from the National Geomagnetic Data Center (NGDC) to form a high-resolution model of the field. Spherical harmonic analysis of the dataset yielded Gauss coefficients up to degree $n = 720$. The power spectrum of the NGDC-720 model field – a plot of the energy density associated with each degree n – is shown in Fig. 11.9. The figure also shows the computed wavelength, or spatial extent, of each part of the spectrum, which is inversely related to the degree n.

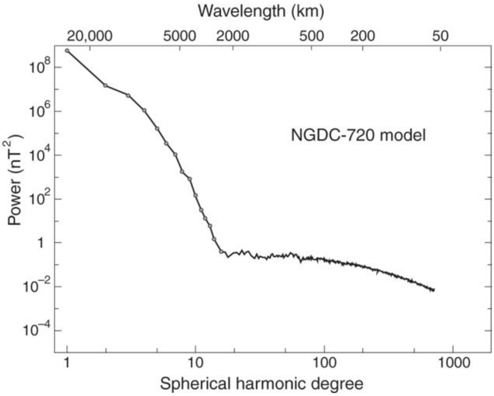


Fig. 11.9 The energy density spectrum of the geomagnetic field derived from a database that combines measurements from the CHAMP satellite with the results of aeromagnetic and marine magnetic surveys.

Source: S. Maus, *Geophysics. J. Int.*, 174, 135–142, 2008.

As described in the interpretation of the source of gravity anomalies (Section 4.2.1), the wavelength of a magnetic anomaly is a measure of the depth of its source. A discontinuity in the slope of the power spectrum at around $n = 15$ separates the sources. The terms $n \leq 14$ dominate the Earth's magnetic field and are due to deep sources in the fluid core. Because the vertical scale in Fig. 11.9 is logarithmic, the terms for $n = 1$, 2, and 3 are much stronger than the rest of the energy spectrum. Each of these parts of the internal field has a different geometry, corresponding to a dipole, quadrupole, and octupole, respectively (Box. 11.3). The terms of degree $n > 15$ describe shorter-wavelength magnetic anomalies, which are associated with the magnetization of the lithosphere. Magnetic field measurements from the SWARM mission have been used to model the lithospheric field in great detail, based on coefficients in the range $15 \leq n \leq 80$ magnetic field.

11.2.4.1 Magnetic Field on the Surface of the Core

The inverse dependence of the field of internal origin (i.e., $n \leq 14$) on radial distance (Eq. 11.6) means that it increases in strength with depth in the Earth. Disregarding electric currents in the mantle or crust as possible sources of the field, it is possible to calculate the magnetic field on the surface of the core. The process, called *downward continuation*, is used extensively in magnetic (and gravity) exploration. It makes use of the $(R/r)^{n+1}$ radial dependence of a term of degree n in the potential. Using 3480 km, the mean radius of the core, for r, and 6371 km for R, a harmonic of degree n in the potential increases by the factor $(1.83)^{n+1}$ when downward continued to the core surface. This results in relative amplification of spherical harmonics with large n (small wavelength), and reveals small-scale features that are not evident at the Earth's surface. Customarily the radial field is computed from the modified potential and mapped on the core surface (Fig. 11.10).

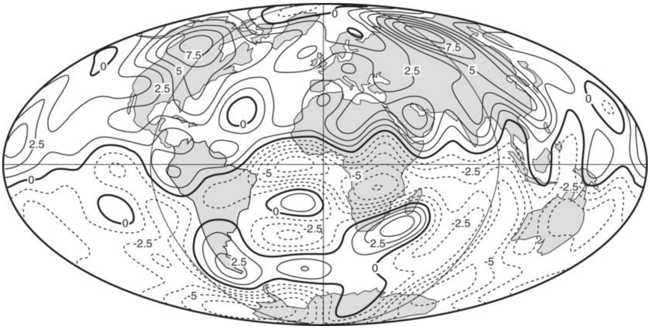


Fig. 11.10 Contours of the radial component of the geomagnetic field on the surface of the core (courtesy of Christopher Finlay, 2018). The numerical values indicate field strength in Gauss units (1 G = 10^5 nT). Solid and dashed contours bound regions where the magnetic field is leaving and entering the core, respectively; the heavy black contours mark null values of the radial field.

Source: International Geomagnetic Reference Field 1990.

The core field has several prominent features. In general, flux enters the core in the northern hemisphere and leaves it in the southern hemisphere, except for a few anomalous regions. At the Earth’s surface the dipole field has a single region of high flux in each hemisphere where field lines come together at each pole. On the core surface the magnetic field shows two regions, referred to as flux lobes, for each pole, aligned at the edge of the tangent cylinder to the inner core. Moreover, in a broad region under the Southern Atlantic Ocean the magnetic flux is entering the core instead of leaving it. This reverse flux patch modifies the field at the Earth’s surface, reducing its intensity over the region and producing the South Atlantic magnetic anomaly.

11.2.4.2 The Geomagnetic Dipole Field

The most important part of the Earth’s magnetic field at the surface of the Earth is the dipole field, given by the Gauss coefficients for which $n = 1$. If we write only the first term of Eq. (11.6) we get the potential

$$W = R^3 g_1^0 \frac{\cos\theta}{r^2}.$$
(11.7)

Note that the spatial variation of this potential depends on $(\cos\theta/r^2)$ in the same way as the potential of a dipole, which was found (Eq. 11.1) to be

$$W = \frac{\mu_0 m}{4\pi}\frac{\cos\theta}{r^2}.$$
(11.8)

Comparison of the coefficients of $(\cos\theta / r^2)$ in Eqs. (11.7) and (11.8) gives the dipole moment of the Earth's axial dipole in terms of the first Gauss coefficient:

$$m = \frac{4\pi}{\mu_0} R^3 g_1^0.$$
(11.9)

The term g_1^0 is the strongest component of the field. It describes a magnetic dipole at the center of the Earth and aligned with the Earth's rotation axis. This is called the *geocentric axial magnetic dipole*.

The magnetic field **B** of a dipole is symmetrical about the axis of the dipole. At any point at distance r from the center of a dipole with moment m on a radius that makes an angle θ to the dipole axis the field of the dipole has a radial component B_r and a tangential component B_θ, which can be obtained by differentiating the potential with respect to r and θ, respectively:

$$B_r = -\frac{\partial W}{\partial r} = \frac{\mu_0 m}{4\pi}\frac{2\cos\theta}{r^3},$$
(11.10)

$$B_\theta = -\frac{1}{r}\frac{\partial W}{\partial \theta} = \frac{\mu_0 m}{4\pi}\frac{\sin\theta}{r^3}.$$
(11.11)

Note that B_r vanishes at the equator ($\theta = 90°$) and the field is horizontal. Using m from Eq. (11.9) and $\sin\theta = 1$ in Eq. (11.11), we find that the horizontal equatorial field B_θ at Earth's surface ($r = R$) is equal to g_1^0. At a point (r, θ) on the surface of a uniformly magnetized sphere the magnetic field line is inclined to the surface at an angle I, which is given by

$$\tan I = \frac{B_r}{B_\theta} = 2\cot\theta = 2\tan\lambda.$$
(11.12)

The angle I is called the *inclination* of the field, and θ is the angular distance (or *polar angle*) of the point of observation from the magnetic axis. The polar angle is the complement of the magnetic latitude, λ (i.e., $\theta = 90° - \lambda$). Eq. (11.12) has an important application in paleomagnetism, as will be seen later.

The terms g_1^1 and h_1^1 are the next strongest in the potential expansion. They describe contributions to the potential from additional dipoles with their axes in the equatorial plane. The total dipole moment of the Earth is then obtained from the vector sum of all three components:

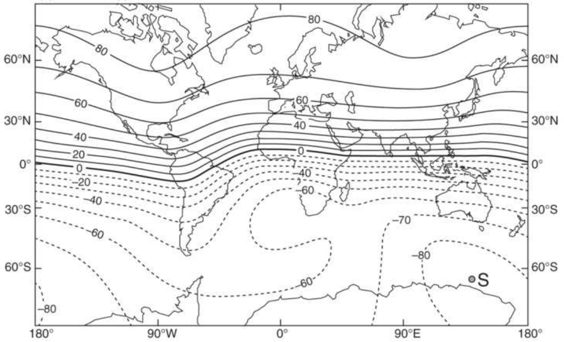
$$m = \frac{4\pi}{\mu_0}R^3\sqrt{(g_1^0)^2 + (g_1^1)^2 + (h_1^1)^2}.$$
(11.13)

Analysis of the geomagnetic field for the year 2015 gave the following values for the dipole coefficients: $g_1^0 = -29{,}442.0$ nT; $g_1^1 = -1501.0$ nT; $h_1^1 = 4797.1$ nT. The Earth's dipole magnetic moment obtained by inserting these values in Eq. (11.13) is $m = 7.724 \times 10^{22}$

A m^2. Note that the sign of g_1^0 is negative. This means that the axial dipole points opposite to the direction of rotation. Taken together, the three dipole components describe a geocentric dipole inclined at about 9.7° to the Earth's rotation axis. This tilted geocentric dipole accounts for more than 90% of the geomagnetic field at the Earth's surface. Its axis cuts the surface at the north and south *geomagnetic poles*. For epoch 2015 the respective poles were located at 80.37°N, 72.63°W (i.e., 287.4°E) and 80.37°S, 107.37°E. The geomagnetic poles are antipodal (i.e., exactly opposite) to each other.

The *magnetic dip poles* of the Earth are defined as the locations where the inclination of the magnetic field is $\pm 90°$ (i.e., where the field is vertically upward or downward). An isoclinal map (showing contours of constant inclination) for the year 2015 shows that the location of the north magnetic dip pole was at 86.29°N, 160.06°W (199.94°E), while the south magnetic pole was at 64.28°S, 136.59°E (Fig. 11.11a). These poles are not exactly opposite one another. The magnetic equator is where the inclination is zero, i.e., the field is horizontal. Because of the tilt of the dipole axis the magnetic equator is not parallel to the geographic equator. The discrepancy between the magnetic dip poles and the geomagnetic poles arises because the terrestrial magnetic field is more complex than that of a perfect dipole.

(a) Inclination (°)


60°N
30°N
0°
30°S
60°S
180°
90°W
0°
90°E
180°
80
80
60
60
40
40
20
0
0
20
40
40
60
60
−70
−80
−80
S

(b) Total intensity (µT)

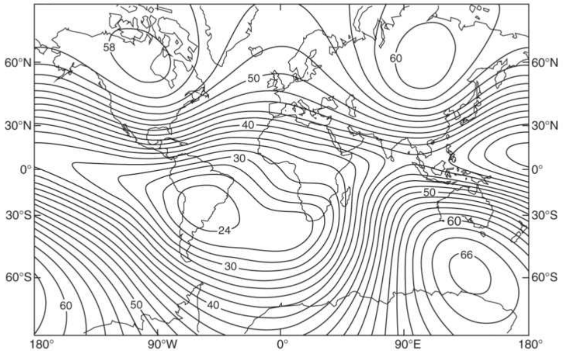

60°N
30°N
0°
30°S
60°S
180°
90°W
0°
90°E
180°
58
50
60
40
30
24
30
50
60
66
60
50
40

Fig. 11.11 Maps of (a) the inclination and (b) the total intensity in µT of the International Geomagnetic Reference Field for the year 201.

Source: E. Thebault et al., International Geomagnetic Reference Field: the 12th generation, *Earth, Planets and Space* 67:79, 2015.

The intensity of the geomagnetic field is generally stronger in high latitudes than near the equator (Fig. 11.11b). The intensity is especially low over the South Atlantic, where it is about 20 µT weaker than expected. Because the geomagnetic field shields the Earth from cosmic radiation and charged particles of the solar wind, this protection is less effective over this region. Orbiting satellites note increased impacts of extra-terrestrial particles over this region. The feature poses hazards for astronauts in low-orbiting spacecraft and also for pilots and passengers in high-altitude aircraft that pass through the region. The enhanced particle flux can interfere with their on-board computers, communications, and guidance systems.

11.2.4.3 The Geomagnetic Non-Dipole Field

The part of the field of internal origin, obtained by subtracting the field of the inclined geocentric dipole from the total field, is collectively called the *non-dipole field*. It amounts to about 5% of the total field. A map of the non-dipole field consists of a system of irregularly sized, long-wavelength magnetic anomalies (Fig. 11.12). To describe this field requires all the terms in the potential expansion of degree $n \geq 2$ in Eq. (11.6).

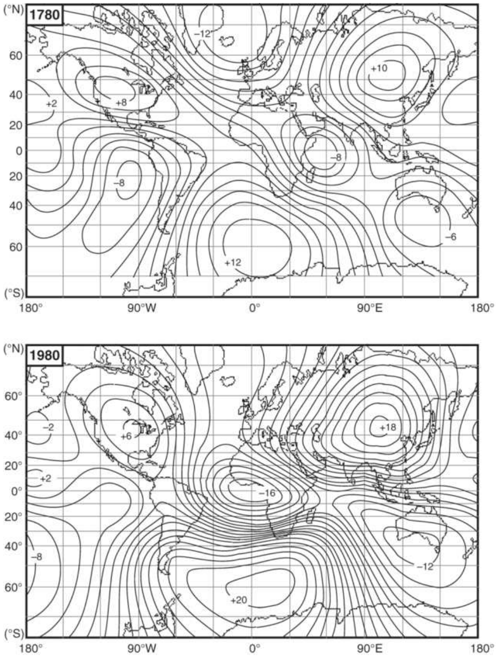

(°N)
1780
60
40
20
0
20
40
60
(°S)
180°
90°W
0°
90°E
180°
−12
+10
+2
+8
−8
−8
+12
−6
(°N)
1980
60°
40°
20°
0°
20°
40°
60°
(°S)
180°
90°W
0°
90°E
180°
−2
+6
+18
+2
−16
−8
−12
+20

Fig. 11.12 The vertical component of the non-dipole magnetic field for the years AD 1780.

Sources: T. Yukutake and H. Tachinaka, The non-dipole part of the Earth's magnetic field, *Bull. Earthq. Res. Inst.* 46, 1027–1062, 1968; C.E. Barton, Geomagnetic secular variation: direction and intensity, in *The Encyclopedia of Solid Earth Geophysics*, Ed. D.E. James, pp. 560–577, New York: Van Nostrand Reinhold, 1989.

11.2.5 Secular Variation

At any particular place on the Earth the geomagnetic field is not constant in time. When the Gauss coefficients of the internal field are compared from one epoch to another, slow but significant changes in their values are observed. The slow changes of the field only become appreciable over decades or centuries of observation and so they are called *secular variations* (from the Latin word *saeculum* for a long age). They are manifest as changes in the Gauss coefficients with time, and affect descriptions of both the dipole and non-dipole components of the field.

11.2.5.1 Secular Variation of the Dipole Field

Direct measurements of the geomagnetic field since the 1840s have been augmented by the analysis of more than 83,000 observations made by mariners between about AD 1590 and AD 1990 (Jackson et al., 2000). The results document the secular variations of intensity and direction of the geomagnetic field during these four centuries. Calculations of the Gauss coefficients show a near-linear decay of the strength of the dipole moment at a rate of about 3.2% per century (Fig. 11.13a). At the start of the twentieth century the decay became even faster and has since averaged about 6.4% per century. If it continues at the same almost linear rate, the field intensity would reach zero in about another 1600 years. The cause of the change in intensity decay rate is not known; it may simply be part of a longer-term fluctuation. However, another possibility is that the dipole moment may be decreasing preparatory to the next reversal of geomagnetic field polarity.

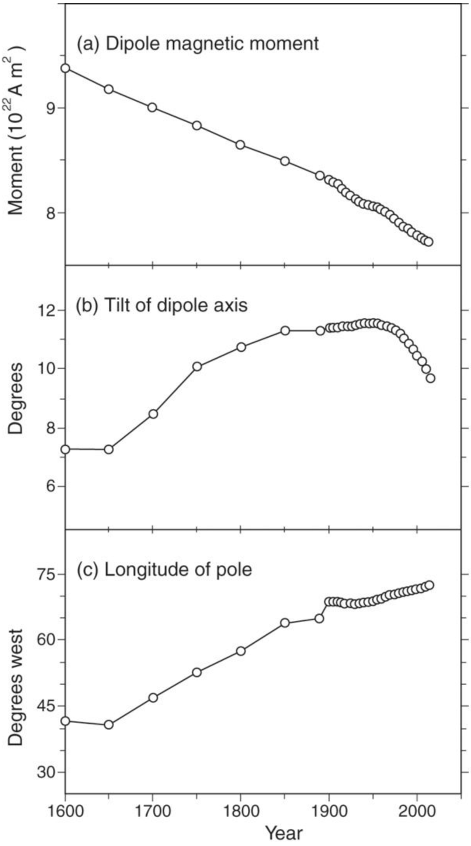


Fig. 11.13 Secular variations of the tilted geomagnetic centered dipole from AD 1600 to AD 2015. (a) Decrease of dipole moment; (b) tilt of the dipole axis relative to the rotation axis; and (c) westward increase in longitude of the geomagnetic poles. Data before 1900 are from Barraclough (1974); data for 1990–2015 are from IRGF analyses (Thebault et al, 2015).

The position of the dipole axis also shows secular variation. The changes can be traced by plotting the colatitude (the angle between the dipole axis and the rotation axis) and longitude of the geomagnetic pole as a function of time. Data are only sufficiently abundant for spherical harmonic analysis since the early nineteenth century. Less reliable data, enlarged by archeomagnetic results (Section 12.2.2.1), allow estimates of the secular variation of the dipole axis since the middle of the sixteenth century. The archeomagnetic data suggest that in the sixteenth century the dipole axis was tilted at only about 3° to the rotation axis; a gradual increase in tilt at about 1.5° per century took place between the sixteenth and nineteenth centuries. From 1850 to 1980 the dipole axis maintained an almost constant tilt of about 11–12° to the rotation axis, but the tilt has decreased by almost 2° since the late 1970s (Fig. 11.13b).

For the past 400 years the longitude of the geomagnetic pole has drifted steadily westward (Fig. 11.13c). Before the nineteenth century the pole moved westward at about 11° per century, corresponding to a pseudo-period of 3300 years for a complete circle about the geographic pole. However, since the early nineteenth century the westward motion of the pole has been slower, at an average rate of 5° per century, which corresponds to a pseudo-period of around 7000 years.

11.2.5.2 Secular Variation of the Non-Dipole Field

The slowly changing values of the Gauss coefficients of degree $n \geq 2$ result in secular variation of the non-dipole field. These are manifest in the differences between maps of the non-dipole field for different epochs (Fig. 11.12). Some anomalies (e.g., over Mongolia, the South Atlantic, and North America) appear to be stationary but change in intensity. Other anomalies (e.g., over Africa) slowly change position with time.

Although some foci may have a north–south component of motion, the most striking feature of the secular variation of the recent non-dipole field is a slow westward drift. Analyses indicate that the secular variation is very low over the Pacific Ocean, so that the magnetic field is more closely dipolar over this region. The evidence for westward drift appears to occur over the rest of the globe. It is superposed on the westward motion of the dipole, but can be separated readily by spherical harmonic analysis. The rate of drift of the non-dipole field can be estimated from longitudinal changes in selected features plotted for different epochs. The mean rate of westward drift in the first half of the twentieth century was estimated to be 0.18 deg yr^{-1}, corresponding to a time of about 2000 years for a full global circuit. However, some foci drift much faster than the average rate and there is little support for the view that drifting features last long enough to complete a global circuit. There is, in fact, evidence that eastward drift of the field may also have occurred. Archeomagnetic data from Europe, for example, document westward drift from AD 1400 to AD 1800, but this was preceded by eastward drift between AD 800 and AD 1400. It is possible to over-interpret these observations: the database supporting westward and eastward drift represents a very short interval of time, a fraction of that required for a global circuit.

Modern analyses of the geomagnetic field allow its representation in spherical harmonic coordinates to a high degree and order. When the dipole component is removed, the variation of the remaining, non-dipole field around a chosen circle of latitude can be computed and filtered to remove the long-period part of the signal. Repetition of this process for the successive versions of the reference fields reveals the variation of the non-dipole field as a function of longitude and time. The results for the years 1650–1950 are shown in Fig. 11.14. They form a banded structure of successively positive and negative anomalies around a circle of latitude (in this case, the equator). The bands are tilted as a result of the westward drift. The drift speed can be estimated from the slope of the bands, which gives a mean equatorial drift rate of 0.27 deg yr^{-1} or 17 km yr^{-1}.

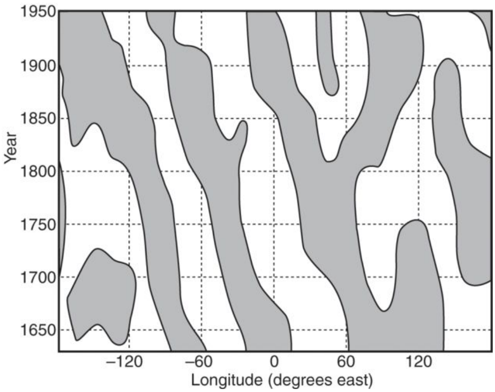


Fig. 11.14 Time–longitude plot of the radial magnetic field at the surface of the core from 1590 to 1990. The time-averaged axially symmetric field has been removed. For each year the radial intensity at the equator is plotted as a function of longitude. The plot shows the zero contours and positive anomalies are shaded.

Source: C.F. Finlay and A. Jackson, GEOPHYSICS: Equatorially Dominated Magnetic Field Change at the Surface of Earth's Core *Science*, 300, 2084–2086, 2003.

Westward drift is an important feature of the secular variation of the historic field. Its origin has been investigated by modeling the physical mechanisms that generate the main geomagnetic field in the Earth's core. One hypothesis associates the westward drift with an oscillatory behavior of the rotating core fluid, although the mechanism that would cause the oscillations is not understood. Another theory is that the outermost layers of the fluid core rotate more slowly than deeper layers, dragging the magnetic field with it and resulting in a westward motion of the non-dipole field relative to the mantle. This has been modeled by downward continuing the secular variation of the 2003 reference field to the core's surface, and computing the tangential fluid flow vectors at the core surface (Fig. 11.15). The flow is dominated by equatorial westward flow and a gyre under the Pacific Ocean. According to the frozen-flux hypothesis (Section 11.2.6.2), the high electrical conductivity of the core fluid causes magnetic field lines to be dragged along by the fluid flow. The estimated flow speeds at the surface of the core are thus of the order of 20 km yr^{-1}.

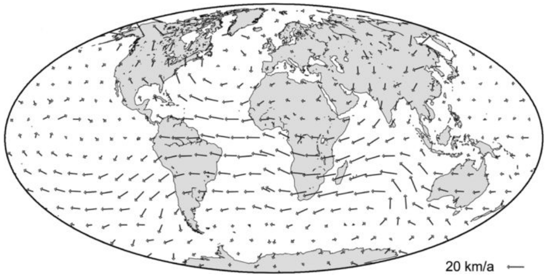


Fig. 11.15 Computed velocity vectors for fluid motion on the surface of the core.

Source: S. Maus, L. Silva, and G. Hulot, J. Can core–surface flow models be used to improve the forecast of the Earth's main magnetic field, *Geophys. Res.*, 113, B08102, 2008

The geomagnetic secular variation occasionally makes an abrupt change, which disrupts a previously monotonous behavior. A plot of the annual change in the eastward component of the field (Fig. 11.16) is characterized by a succession of nearly linear segments separated by abrupt changes of slope known as *geomagnetic impulses* (more commonly called *"geomagnetic jerks"*). They may be caused by sudden, as yet unexplained, changes in the fluid flow near the surface of the outer core.

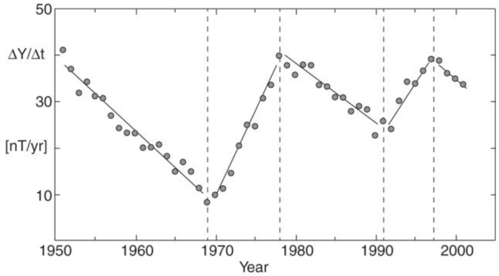


Fig. 11.16 Secular variation of the geomagnetic field observed at Niemegk observatory, Germany. Points represent the differences between annual averages of the eastward (Y-) component of the field. Linear segments are interrupted by sudden directional changes, called geomagnetic jerks, in 1969, 1978, 1991, and 1997 (marked by vertical dashed lines).

Source: Bloxham et al., The origin of geomagnetic jerks, *Nature,* 420, 65–68, 2002

11.2.6 Origin of the Internal Field

Analysis of the power spectrum of the geomagnetic field (Fig. 11.9) indicates that the main field is produced in the fluid outer core of the Earth. The composition of the fluid core has been estimated from seismic and geochemical data. The major constituent is liquid iron, with smaller amounts of other, less-dense elements. Geochemical analyses of iron meteorites suggest that the core composition may have a few percent of nickel, while shock-wave experiments require 6–10% of non-metallic light elements such as silica, sulfur, or oxygen. The solid inner core is inferred from seismic and shock-wave data to consist of almost pure iron.

For the generation of the magnetic field the important parameters of the core are its temperature, viscosity, and electrical conductivity. Temperature is known very poorly inside the Earth, but probably exceeds 3000°C in the liquid core. The electrical conductivity of iron at the high temperatures and pressures in the core is estimated to be around $1.1\text{–}1.2 \times 10^6$ S m^{-1} (Pozzo et al., 2012), which corresponds to a very good electrical conductor. For comparison, the conductivity of carbon (used for many electrical contacts) is 3×10^4 S m^{-1} at 20°C.

11.2.6.1 Magnetostatic and Electromagnetic Models

As observed by Gilbert in 1600, the dipole field of the Earth resembles that of a uniformly magnetized sphere. However, permanent magnetization is an inadequate explanation for the geomagnetic dipole. The mean magnetization of the Earth would need to be many times greater than that of the most strongly magnetic crustal rocks. The Curie point of the most important minerals at atmospheric pressure is less than 700°C, which is reached at depths of about 25 km, so that only a thin outer crustal shell could be permanently magnetized. The necessary magnetization of this shell is even greater than values observed in crustal rocks. Moreover, a magnetostatic origin cannot account for the temporal changes observed in the internal field, such as its secular variation.

The main magnetic field of the Earth is thought to be produced by electrical currents in the conductive core. Although the core is a good conductor, an electrical current system in the core continually loses energy through ohmic dissipation. The lost electrical energy is converted to heat and contributes to the thermal balance of the core. The equations of electromagnetism applied to the core show that an electrical current in the core would decay to zero in around 10,000 years unless it is sustained. Paleomagnetic evidence in the form of coherently magnetized rocks supports the existence of a geomagnetic field since Precambrian time, i.e., for about 3000 Myr. This implies that it must be continuously maintained or regenerated. The driving action for the main field is called the *dynamo process*, by analogy to the production of electrical power in a conductor that rotates in a magnetic field.

11.2.6.2 The Geomagnetic Dynamo

When a charged particle moves through a magnetic field, it experiences a deflecting electrical field (called the Lorentz field) proportional to the magnetic flux density B and the particle velocity v, and acting in the direction normal to both B and v. The Lorentz field acts as an additional source of electrical current in the core. Its strength is dependent on the velocity of motion of the conducting fluid relative to the magnetic field lines. When this term is included in the Maxwell electromagnetic equations, a *magnetohydrodynamic* equation relating the magnetic field $\mathbf{B}$ to the fluid flow $\mathbf{v}$ and conductivity σ in the core is obtained. It is written as

$$\frac{\partial \mathbf{B}}{\partial t} = \frac{1}{\mu_0 \sigma} \nabla^2 \mathbf{B} + \nabla \times (\mathbf{v} \times \mathbf{B}). \tag{11.14}$$

This vector equation, although complicated, has immediate consequences. The left side gives the rate of change of magnetic flux in the core; it is determined by two terms on the right side. The first is inversely dependent on the electrical conductivity, and determines the decay of the field in the absence of a driving potential; the better the conductor, the smaller is this *diffusion term*. The second, *dynamo term* depends on the Lorentz electrical field, which is determined by the velocity field of the fluid motions in the core. The conductivity of the outer core (–10^6 Ω^{-1} m^{-1}) is high and for a fluid velocity above about 1 mm s^{-1} the dynamo term greatly exceeds the diffusion term. Under these conditions, the lines of magnetic flux in the core are dragged along by the fluid flow. This concept is called the *frozen-flux theorem*, and it is fundamental to dynamo theory. The diffusion term is only zero if the electrical conductivity is infinite. There is probably some diffusion of the field through the fluid, because it is not a perfect conductor. However, the frozen-flux theorem appears to approximate well the conditions in the fluid outer core.

The derivation of a solution of the dynamo theory is difficult. In addition to Maxwell's equations with the addition of a term for the Lorentz field, the Navier–Stokes equation for the fluid flow, Poisson's equation for the gravitational potential, and the generalized equation of heat transfer must be simultaneously satisfied. The fluid flow consists of a radial component and a rotational component. The energy for the radial flow comes from two sources. The slow cooling of the Earth produces a temperature gradient in the core, which results in thermally driven convection in the iron-rich fluid of the outer core. This is augmented by latent heat released at the boundary between the inner and outer cores as the inner core solidifies. The solidification of the pure iron inner core depletes the fluid of the outer core of its heaviest component. The remaining lighter elements rise through the liquid outer core, causing buoyancy-driven convection.

The rotational component of the fluid flow is the result of a radial velocity gradient in the liquid core, with inner layers rotating faster than outer layers. The relative rotation of the conducting fluid drags magnetic field lines around the rotational axis to form ring-like, *toroidal* configurations. The toroidal field lines are parallel to the flow and therefore to the surface of the core. This means that the toroidal fields are confined to the core and cannot be measured; their strengths and configurations must be estimated from models. Their interactions with the upwelling and descending branches of convective currents create electrical current systems that produce poloidal magnetic fields. These, in their turn, escape from the core and can be measured at the surface of the Earth. The fluid motions are subject to the effects of Coriolis forces, which prove to be strong enough to dominate the resultant flow patterns.

Studies of the paleomagnetic properties of some Precambrian rocks (Section 12.2.5) that date from 3.45 Ga have documented that they possess remanent magnetizations, which implies that the Earth's magnetic field must be at least this old. However, the solid inner core is believed to be much younger. Thermal models suggest that it dates from 0.5–2 Ga. The compositional component of core convection can only have existed since nucleation of the inner core, so thermal convection alone must have powered the geomagnetic dynamo before then, perhaps in conjunction with a different kind of compositional convection. This matter is as yet unresolved. The growth of the inner core and the age of the geomagnetic field are important themes of current research into the early evolution of the Earth.

11.2.6.3 Simulation of the Geodynamo

Important advances in understanding the geomagnetic dynamo (or *geodynamo*) have been made by simulating the related core processes with supercomputers. In 1995, G. A. Glatzmaier and P. H. Roberts presented a numerical model for the generation of a magnetic field, assuming a hot convective fluid outer core surrounding a solid inner core, with rotation rate akin to that of the Earth. The heat flow, electrical conductivity, and other material properties were made as similar as possible to those of the Earth's core. The simulated dynamo had dominantly dipole character and intensity similar to the Earth's, and it exhibited a westward drift of the non-dipole field comparable to that measured at the Earth's surface. It also reversed polarity spontaneously, with long periods of constant polarity between short polarity transitions, as is the case for the history of geomagnetic reversals in the last 160 Myr (Section 12.3). During a polarity reversal, the field intensity decreased by an order of magnitude, as is also observed in paleomagnetic studies (see Fig. 12.27).

The simulations showed that the ability to reverse polarity was increased when the heat flow across the core–mantle boundary was non-uniform, as in the Earth, showing that thermal conditions in the lower mantle influence the formation of the magnetic field in the fluid core. The solid inner core evidently plays an important role in the reversal process. Magnetic fields in the outer core can change quickly, accompanying convection, and may act to initiate reversals. However, magnetic fields in the solid inner core change more slowly by diffusion and thus the inner core may act to stabilize the field against reversals so that they occur only occasionally.

11.3 Magnetic Fields of the Sun, Moon, and Planets

11.3.1 Magnetic Fields of the Sun and Moon

Our knowledge of the magnetic fields of the Sun and planets derives from two types of observation. Indirect observations utilize spectroscopic effects. All atoms emit energy related to the orbital and spin motions. The energy is quantized so that the atom possesses a characteristic spectrum of energy levels. The lowest of these is called the ground state. When an atom happens to have been excited to a higher energy level, it is unstable and eventually returns to the ground state. In the process, the energy corresponding to the difference between the elevated and ground states is emitted as light. For example, atomic hydrogen gas in the Sun and galaxies emits radiation at 1420 MHz, with corresponding wavelength of 21 cm. This frequency is in the microwave range and can be detected by radio telescopes. If the hydrogen gas is in motion, the frequency is shifted by the Doppler effect, from which the velocity of the motion can be deduced.

In the presence of a magnetic field a spectral line can split into several lines. This “hyperfine splitting” is called the *Zeeman effect*. When a hydrogen atom is in the ground state its hyperfine structure has only one line. However, if the atom is in a magnetic field, it acquires additional energy from the interactions between the atom and the magnetic field and hence it can exist in several energy states. As a result, the spectral lines of energy emitted by the atom are split into closely spaced lines, representing transitions between the different energy states. The energy differences between these states are dependent on the strength of the magnetic field, which can be estimated from the observations.

Direct observations of extra-terrestrial magnetic fields have been carried out since the 1960s by space probes. Magnetometers mounted in these spacecraft have recorded directly the intensity of the interplanetary magnetic field as well as the magnetic fields around several planets. The manned Apollo missions to the Moon resulted in a large amount of data being obtained from the orbiting spacecraft. The materials collected on the Moon's surface and brought back to Earth by the astronauts have provided valuable information about lunar magnetic properties.

11.3.1.1 Magnetic Field of the Sun

The Sun has nearly 99.9% of the mass of the solar system. About 99% of this mass is concentrated in a massive central core that reaches out to about 80% of the Sun's radius. The remainder of the Sun consists of an outer conducting shell that contains only 1% of the Sun's mass but has a thickness equivalent to about 20% of the solar radius. Thermonuclear conversion of hydrogen into helium in the dense core produces temperatures of the order of 15,000,000 K. The visible solar disk is called the *photosphere*. Its diameter is about 2,240,000 km (about 175 times the diameter of the Earth) and its surface temperature is about 6000 K. The lower solar atmosphere is called the *chromosphere*; the outer atmosphere is called the *corona*. The chromosphere includes spike-like emissions of hydrogen gas called solar prominences that sometimes can reach far out into the corona.

The heat from the core is radiated out to the outer conducting shell, where it sets up convection currents. Some of these convection currents are small-scale (about 1000 km across) and last only a few minutes; others are larger-scale (30,000 km across) and persist for about a (terrestrial) day. The convection is affected by the Sun's rotation and is turbulent.

The rotation of the Sun has been estimated spectroscopically by measuring the Doppler shift of spectral lines. Independent estimates come from observing the motion of solar features like sunspots, etc. The rotational axis is tilted slightly at 7° to the pole to the ecliptic. Near the Sun's equator the rotational period is about 25 Earth days; the polar regions rotate more slowly with a period of 35–40 days. The Sun's core may rotate more rapidly than the outer regions. Turbulent convection and velocity shear in the outer conducting shell are conducive to a dynamo origin for the Sun's magnetic field. The surface field is dipolar in higher latitudes but has a more complicated pattern in equatorial latitudes.

Temperatures in the outer solar atmosphere (corona) are very high, around 1,000,000 K. The constituent particles achieve velocities in excess of the escape velocity of the Sun's gravitational field. The supersonic stream of mainly protons (H^+ ions), electrons and α-particles (He^{2+} ions) escaping from the Sun forms the solar wind. The flow of electric charge produces an interplanetary magnetic field (IMF) of varying intensity. At the distance of the Earth from the Sun the IMF measures about 6 nT.

Sunspots have been observed since the invention of the telescope (*c.* AD 1610). A sunspot is a dark fleck on the surface of the Sun, measuring roughly 1000 to 100,000 km in diameter. It has a lower temperature than the surrounding photosphere and represents a strong disturbance extending far into the Sun's interior. Sunspots last for several days or weeks and move with the Sun's rotation, providing a means of estimating the rotational speed. The frequency of sunspots changes cyclically with a period of 11 years.

Intense magnetic fields are associated with the sunspots. These often occur in unequally sized pairs of opposite polarity. The predominating magnetic polarity of sunspots changes from one period of maximum sunspot activity to the next, implying that the period is in fact 22 years. The magnetic field of each sunspot is toroidal. It can be imagined to resemble a vortex, or tornado, in which the magnetic field lines leave the solar surface in one sunspot and return to it in the other member of the pair. The polarity of the Sun's dipole field also reverses with the change of polarity of the sunspots, indicating that the features are related.

Associated with the sunspots and their strong magnetic fields are emissions of hydrogen gas, called *solar flares*. The charged particles ejected in the flares contribute to the solar wind, which consequently transfers the sunspot cyclicity to fluctuations in the Earth's magnetic field of external origin and to ancillary terrestrial phenomena such as magnetic storms, brilliant aurora, and interference with radio transmissions.

11.3.1.2 Magnetic Field of the Moon

Classical hypotheses for the origin of the Moon – rotational fission, capture, or binary accretion – have been superseded by the Giant Impact hypothesis. According to this model, the Earth experienced a catastrophic collision with a Mars-sized protoplanet early in its development. Some debris from the collision remained in orbit around the Earth, where it re-accumulated as the Moon about 4.5 Gyr ago. This origin accounts for why the Earth and Moon have the same oxygen isotope composition. At the time of the Giant Impact the Earth's mantle and core had already differentiated and dense iron had settled into the core. This may also have taken place in the impactor so the Moon formed from the iron-depleted rocky mantles of the impacting bodies. The Moon has only a small core, whereas if it had originated like the other planets its core would be proportionately larger.

The early Moon was probably covered by an ocean of molten magma at least a hundred kilometers thick, the relicts of which are now the major constituents of the lunar highlands. After cooling and solidification the lunar surface was bombarded by planetesimals and meteoroids until about 4 Ga ago. The huge craters left by the impacts were later filled by molten basalt to form the lunar maria. Rock samples recovered from the maria in the manned Apollo missions have been dated radiometrically at 3.1–3.9 Ga. The volcanism may have continued after this time but probably ceased about 2.5 Gyr ago. Since then the lunar surface has been pulverized by the constant bombardment by meteorites, micrometeorites, and elementary particles of the solar wind and cosmic radiation. Except for the lunar highlands the surface is now covered by a layer of shocked debris several meters thick called the lunar *regolith*.

Our knowledge of lunar magnetism derives from magnetic field measurements made from orbiting spacecraft, magnetometers set up on the Moon, and rock samples collected from the lunar surface and brought back to Earth. Measurements of the lunar surface field from orbiting spacecraft have been made by on-board magnetometers and by the electron reflection method, the principle of which is basically the same as used to explain the origin of the Van Allen belts in the Earth's magnetosphere (Section 11.2.3.1). Electrons rain abundantly upon the lunar surface, in part from the solar wind and in part from the charges trapped in the Earth's magnetotail, which is crossed by the monthly orbit of the Moon about the Earth. In the absence of a lunar magnetic field the electrons would be absorbed by the lunar surface. As in the Earth's magnetosphere an electron incident on the Moon is forced by the Lorentz force to spiral about a magnetic field line (see Fig. 11.6). As the electron approaches the lunar surface the magnetic field strength increases and the pitch of the helix decreases to zero. At the point of closest approach (also called the *mirroring point*), the electron motion is completely rotational about the field line. The electron path then spirals with increasing pitch back along the field line, away from the Moon. The effectiveness of the electron reflection is directional; for a given electron velocity there is a critical angle of incidence below which reflection ceases. By counting the number of electrons with a known energy that are reflected past a satellite in known directions the surface field B_s can be estimated from the field B_0 measured at the altitude of the satellite.

The lunar magnetic field was surveyed extensively in 1998 by the Lunar Prospector orbiting spacecraft. The measurements show that the Moon currently has no detectable dipole moment. It has only a very weak non-dipolar magnetic field due to local magnetization of crustal rocks. Estimates of surface anomalies suggest that the largest are of the order of 100 nT. From records of meteoritic impacts and moonquakes made by seismometers left on the Moon for several years, it has been inferred that, if the Moon has a core (necessary for a lunar dynamo), it must be smaller that 400–500 km in radius, representing less than 2% of the Moon’s volume and 1–3% of its mass. In contrast, the Earth’s core occupies about 16% of the planet’s volume and has about 33% of its mass. Lunar Prospector detected magnetic field changes due to electrical currents induced in the Moon as it traversed the stretched-out tail of Earth’s magnetosphere They suggest an even smaller, iron-rich metallic core, with radius 340 ± 90 km (Hood et al., 1999).

Although the Moon now has no global dipole magnetic field, samples of lunar rocks recovered in the manned Apollo missions possessed quite strong natural remanent magnetizations. Rock magnetic studies on Apollo samples with radiometric ages of ∼3.6–3.9 Ga suggest that they were magnetized in fields of the order of 10–100 μT (0.1–1 gauss), much stronger than present fields on the Moon. As yet, the ancient lunar magnetic field is not well understood. The interior of the Moon may have acquired a primordial remanent magnetization in the external field of the Sun or Earth. In turn, this may have provided the fields for acquisition of the observed crustal magnetizations. The Moon presently has a small, solid metallic inner core, surrounded by a thin fluid outer core, but the core heat flux is too low to have powered an internal dynamo at the time of formation of the Apollo samples. However, models suggest that a thin layer of the Moon’s mantle adjacent to the core may have first blanketed the core, then later provided enough radioactive heating to assist dynamo activity that persisted for a limited period during a “magnetic era” about 0.5–1.0 Ga after the Moon was formed (Stegman et al., 2003).

11.3.2 Magnetic Fields of the Planets

Jupiter and Saturn, like the Earth, have magnetic fields that are strong enough to trap charged particles. The motions of these charged particles generate electromagnetic radiation which is detectable as radio waves far from the planet. The magnetic field of Jupiter was first detected in this way. Moreover, the radio emissions are modulated by the rotation of the planet. Analysis of the periodicity of their modulated radio emissions provide the best estimates of the rotational rates of Jupiter and Saturn.

Most data concerning the magnetic fields of the planets (Table 11.1) have been obtained with flux-gate magnetometers (Section 11.4.2.1) on-board passing or orbiting spacecraft. When the spacecraft traverses the magnetosphere of a planet, the magnetometer registers the passage through the bow shock and magnetopause (Fig. 11.17). A bow shock results from the supersonic collision of the solar wind with the atmosphere of a planet, just as it does for the Earth (see Fig. 11.4). Counters of energetic particles register a sudden increase in frequency and the magnetometer shows a change in magnetic field strength during passage through the bow shock into the magnetosheath. The magnetopause is where the pressure exerted by the moving solar wind is equal to the pressure on the plasma from the planetary magnetic field. When the spacecraft leaves the magnetosheath and crosses the magnetopause it enters a region that is largely shielded from the solar wind by the magnetic field of the planet. The existence of a bow shock is evidence for a planetary atmosphere, while the magnetopause is evidence that the planet has a magnetic field.

Table 11.1 Magnetic characteristics of the planets

Venus and Mars do not have a global magnetic field. Venus and Uranus have retrograde rotations. The dipole magnetic moment ($g_1^0 R_P^3$) is calculated relative to the Earth's ($m_E = 7.654 \times 10^{15}$ T

m^3).

Planet	Equatorial radius R_p (km)	Axial rotation period (days)	Dipole magnetic moment (m_E)	Equatorial magnetic field g_1^0 (nT)	Dipole tilt to rotation axis (°)
Mercury	2440	58.65	0.0004	195	~0
Venus	6052	–243.0	–	–	–
Earth	6378	1	1	29,500	9.7
Mars	3396	1.026	–	–	–
Jupiter	71,492	0.414	20,000	410,000	10.3
Saturn	60,268	0.444	600	21,000	<1
Uranus	25,559	–0.718	50	22,800	58.6
Neptune	24,764	0.671	28	14,200	46.9

Source: NASA, planetary fact sheets.

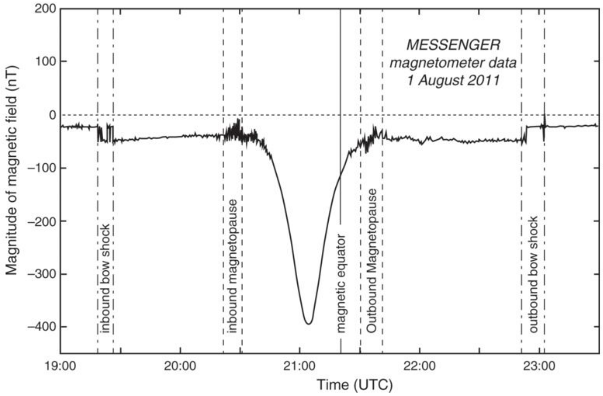


Fig. 11.17 Magnetic field measurements along the path of the MESSENGER spacecraft during an orbit of the planet Mercury.

Data source: Johnson et al., MESSENGER observations of Mercury's magnetic field structure, *J. Geophys. Res.*, E00L14, 2012.

11.3.2.1 Mercury, Venus, and Mars

Mercury was visited by the Mariner 10 spacecraft, which made three passes of the planet in 1974 and 1975. The on-board magnetometer detected a bow shock and a magnetopause, which imply that the planet has a magnetic field. On the first encounter a magnetic field of about 100 nT was measured at an altitude of 700 km (radial distance of 3100 km), which was the point of closest approach. It was not possible to determine definitively the source of the magnetic field.

The MESSENGER mission to Mercury was launched in 2004 and after a complex passage in which it used the gravity fields of Earth, Venus, and Mercury itself during fly-bys of these planets to adjust its velocity, it completed orbital insertion around Mercury in 2011. A vector magnetometer on the spacecraft measured orthogonal components of the magnetic field; the total intensity along an orbital path is illustrated in Fig. 11.17. The measurements made during more than four years of orbiting the planet revealed the character of Mercury's magnetic field.

The planet has a liquid metallic core surrounding a solid inner core, each of which is proportionately larger than the Earth's relative to the planet's radius; it also has a comparatively thick crust and thin mantle. A dipole field is generated in the core and produces a surface magnetic field of 190 nT, which is less than 1% of the average field at the Earth's surface. Moreover, the dipole is closely aligned with the rotation axis and is stronger in the northern hemisphere than in the southern hemisphere of the planet. This is because the center of the dipole is offset northward along the axis by about 20% of the planet's radius, R_m. The weak internal magnetic field is barely able to stand off the solar wind at the subsolar point to form a magnetosphere, which comes close to the planet. The average subsolar distance of the magnetosphere is only 0.43 R_m. The solar wind causes currents in the magnetosphere, which induce external magnetic fields that can be similar to or stronger than the planetary field. The analysis of short-wavelength magnetic anomalies revealed that they were due to remanent magnetization in the planet's crust and that the planet's magnetic field is ancient, possibly dating from 3.7–3.9 Ga.

Venus was investigated by several American and Russian spacecraft in the 1960s. In 1967 the Soviet Venera 4 spacecraft landed instruments by parachute. They observed that the magnetic field was much weaker than on Earth, the surface temperature was estimated to be above 260°C, and an atmospheric pressure of 22 atmospheres ultimately terminated the mission. In the same year, the Mariner 5 detected a bow shock from the collision of the solar wind with the planetary atmosphere. Data from the Pioneer Venus orbiter in 1978 set an upper limit of $10^{-5}\ m_E$ for a planetary dipole moment, which would give a surface equatorial field of <1 nT. The Venus Express orbiter from 2006 to 2014 established that Venus has a magnetosphere of induced origin. This arises because high in the planet's atmosphere ions form an encircling ionosphere, which interacts with the charged particles and magnetic field of the solar wind. This diverts the solar wind around the planet, causing a magnetosphere and magnetotail.

It had been expected that Venus would have an intrinsic magnetic field. The magnetic moment of a dipole is proportional to the equatorial field times the cube of the planetary radius (Eq. 11.9). The strength of a dynamo field is expected to be proportional to the core radius and to the rotation rate. These considerations give a scaling law whereby the dipole magnetic moment of a planet is proportional to its rotation rate and to the fourth power of the core radius. The rotation of Venus is very slow compared to that of the Earth; one sidereal day on Venus lasts 243 Earth days, but the size of the planet is close to that of the Earth. It was therefore expected that Venus might have an internal dynamo with a dipole moment about 0.2% that of the Earth and an equatorial surface field of about 86 nT. The absence of a field of this size suggests that the slow rotation may not provide enough energy for an active dynamo. Convection is another necessary dynamo ingredient, and this may be lacking in the Venusian core. Thermal convection would be absent if the core has a uniform temperature. Also, the planet may not have an inner core, which produces a compositional form of convection on Earth.

Mars was expected to have a magnetic moment between that of Earth and Mercury, because of its size and rotation rate. Mariner 4 in

1965 was the first American spacecraft carrying magnetometers to visit Mars; it detected a bow shock but no conclusive evidence for a magnetopause. In September 1997, the Mars Global Surveyor spacecraft entered orbit around the planet. Until the mission ended in 2006 it mapped the Martian magnetic field from an almost circular orbit about 400 ± 30 km above the planet's surface. The survey measurements have nearly uniform global coverage and show that there is no significant global magnetic field at present. This does not exclude the possible existence of a dynamo-generated global field in the planet's distant past. The presence of crustal magnetization suggests that Mars had an active magnetic dynamo early in its history, but this is thought to have become extinct about four billion years ago.

The measured magnetic field consists of large regional magnetic anomalies, which are attributed to remanent magnetization of the Martian crust. At the survey altitude the crustal magnetic anomalies have amplitudes up to 220 nT, an order of magnitude larger than the crustal anomalies of Earth at that altitude, which attests to the presence of strongly magnetic minerals in the Martian crust. The crustal magnetic field shows some features with circular geometry, some of which may be related to impact processes. However, the most striking anomalies are prominent east–west trending linear features over a region in the planet's southern hemisphere. These sub-parallel lineations are attributed to alternating bands of strongly magnetized crust. The origin of the lineated magnetic anomalies on Mars is not yet understood.

11.3.2.2 The Giant Planets and Pluto

Jupiter possesses a massive magnetic field, which was detected in 1955 by means of the polarized radio emissions associated with it. The Pioneer 10 and 11 spacecraft in 1973–1974 and Voyager 1 and 2 in 1979 established that the planet has a bow shock and magnetopause. From 1995 to 2003, the Galileo mission made extensive surveys of

Jupiter's magnetosphere, and since 2016 the *Juno* spacecraft orbiting the planet has provided a detailed picture of Jupiter's magnetic field.

Spherical harmonic analysis (Box 11.3) of the magnetic measurements from the Juno mission – the same technique as used on Earth to establish the IGRF – was carried out to degree and order 20. A subset of coefficients up to degree and order 10 was established as a model of the Jovian field. The energy spectrum of the field (similar to Fig. 11.9) is entirely attributed to the core, because Jupiter has no crust or mantle. Unlike the Earth's core, the Jovian core has no firm surface. Downward continuation of the energy spectrum to find a surface where each degree is equal defined a hypothetical "surface" of the core at 0.85 Jovian radii (R_J). The radial magnetic field on this surface (Fig. 11.18) is characterized by a strong hemispheric difference, with strong anomalies on one hemisphere and a smoother field on the other. An unusual feature is a reversely magnetized equatorial flux patch near longitude 90° W, in which the magnetic field lines are entering the core instead of leaving it.

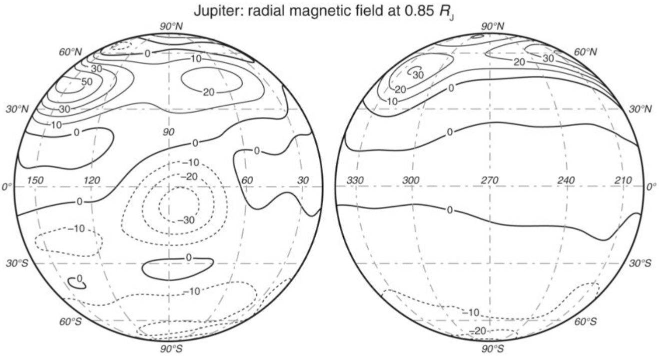


Fig. 11.18 Radial component of the magnetic field on the surface of Jupiter's core, measured by the Juno spacecraft during the initial nine orbits of the planet. Jupiter's magnetic field is much stronger than the Earth's and is charted in Gauss units (1 G = 100 μT).

Source: Connerney et al., A new model of Jupiter's magnetic field from Juno's first nine orbits, *Geophys. Res. Lett.*, 45, 2590–2596, 2018.

The Jovian magnetosphere encounters the solar wind about 5,000,000 km “upwind” from the planet; “downwind,” its magnetotail may at times extend to the orbit of Saturn. Two reasons account for the great size of the magnetosphere compared to that of Earth. First, the solar wind pressure on the Jovian atmosphere is weaker as a result of Jupiter’s greater distance from the Sun. Second, Jupiter’s magnetic field is much stronger than that of the Earth. The Juno mission estimated that the strength of Jupiter’s surface field varies between 200 μT and 2000 μT (2–20 G), arising from a dipole moment of 4.17 G-R_J^3, about 20,000 times stronger than the Earth’s (Table 11.1). The dipole component gives an equatorial magnetic field of more than 410,000 nT at Jupiter’s surface, 14 times the comparable equatorial field on Earth. The quadrupole, octupole, and higher harmonics in Jupiter’s magnetic field are proportionately much larger relative to the dipole field than on Earth. The dipole axis is tilted at 10.3° to the rotation axis, and is also displaced along it by about 12% of Jupiter’s equatorial radius. The magnetic field of Jupiter results from an active dynamo in the metallic hydrogen core of the planet. The core is probably very large, with a radius up to 75% of the planet’s radius. This would explain the high harmonic content of the magnetic field near the planet.

The four Galilean moons orbit within Jupiter’s magnetosphere, which shields them from the solar wind. Io is a volcanically active moon and may be a source of ions injected into the Jovian magnetosphere. Ganymede is almost as large as Mars. It is the only moon in the solar system that has its own magnetosphere. At Ganymede’s equator the moon’s magnetic field is 720 nT, about six times stronger than Jupiter’s field at that location. The magnetic field is probably generated by dynamo action in a metallic core of the moon. The source of the energy needed to power the dynamo is unknown.

Saturn was explored by the spacecraft Pioneer 11 in 1979, Voyager 1 in 1980, and Voyager 2 in 1981. The on-board magnetometers detected a bow shock and a magnetopause. In 2004 the Cassini-Huygens mission explored Saturn's rings and moons and orbited the planet until 2017. Saturn has a magnetosphere that stands off the solar wind about 22 planetary radii from the planet, and a long magnetotail that extends hundreds of planetary radii downwind. It has a strong dipole magnetic moment estimated to be 600 times larger than the Earth's. This provides an equatorial field of about 21,000 nT, slightly weaker than that of the Earth. The magnetic field has a purer dipole character (i.e., the non-dipole components are weaker) than the fields of Jupiter or Earth. The simplest explanation for this feature is that the field is generated by an active dynamo in a conducting core that is small relative to the size of the planet. The axis of the dipole field is inclined at less than 1° to the rotation axis, in contrast to 11.4° on Earth and 9.7° on Jupiter. The center of the dipole is shifted by 3.7% of Saturn's radius along the rotation axis into the northern hemisphere.

Uranus is unusual in that its spin axis has an obliquity of 97.9°. This means that the rotation axis lies very close to the ecliptic plane, and the orbital planes of its satellites are almost orthogonal to the ecliptic plane. Uranus was visited by Voyager 2 in 1986 and 1989. The spacecraft encountered a bow shock and magnetopause and subsequently entered the magnetosphere of the planet. This has an asymmetric geometry, because of the unusual attitudes of the rotation and magnetic dipole axes. The magnetosphere extends for 18 planetary radii (460,000 km) upwind toward the Sun; a long magnetotail extends downwind for many multiples of the planet's radius. Uranus has a dipole moment about 50 times stronger than Earth's, giving a surface field that averages around 23,000 nT, which is comparable to that on Earth. However, the surface field in fact shows a very large variation. There are two reasons for this. First, the axis of the dipole field is tilted at 58.6° to the rotation axis; there is no explanation for this large tilt. Another oddity is that the magnetic field is not centered on the center of the planet, but is displaced by 30% of the planet's radius along the tilted rotation axis. Consequently there is a 10-fold variation in the magnetic field at the planet's surface, from about 10,000 nT to 110,000 nT. The quadrupole component of the field is relatively large compared to the dipole, which suggests that the magnetic field of Uranus is generated at shallow depths within the planet.

Neptune, visited by Voyager 2 in 1989, has a magnetic field with similar characteristics to that of Uranus. Its axis is tilted at 46.9° to the rotation axis and is offset from the planetary center by 55% of Neptune's radius. As in the case of Uranus, the magnetic field has a large quadrupole term compared to the dipole and thus probably originates in the outer layers of the planet, rather than in the deep interior.

Pluto was visited by New Horizons in 2015, but due to weight considerations the spacecraft was not equipped with a magnetometer so no direct magnetic measurements were made. The spacecraft detected a bow shock caused by the interaction of the solar wind with Pluto's thin atmosphere. This creates a boundary layer (referred to as a "Plutopause") around a region of heavy ions in which the solar wind is excluded and which extends downwind from Pluto more than 100 times the planet's radius. Modeling indicates that any global magnetic field would be weaker than 30 nT. The small size, slow rotation rate, and interpreted rocky core make it unlikely that Pluto has a dynamo-generated magnetic field.

11.4 Magnetic Surveying

11.4.1 The Magnetization of the Earth’s Crust

The high-degree terms in the energy density spectrum of the geomagnetic field (Fig. 11.9) are related to the magnetization of crustal rocks. Magnetic investigations can therefore yield important data about geological structures. By analogy with gravity anomalies we define a magnetic anomaly as the difference between the measured (and suitably corrected) magnetic field of the Earth and that which would be expected from the International Geomagnetic Reference Field (Section 11.2.4). The magnetic anomaly results from the contrast in magnetization when rocks with different magnetic properties are adjacent to each other, as, for example, when a strongly magnetic basaltic dike intrudes a less magnetic host rock. The stray magnetic fields surrounding the dike disturb the geomagnetic field locally and can be measured with sensitive instruments called magnetometers.

The Earth’s magnetic field aligns the magnetic moments of mineral grains in a rock and produces an induced magnetization ($\mathbf{M}_i$) that is proportional to the strength of the field, the proportionality constant being the magnetic susceptibility, which can have a wide range of values in rocks (Fig. 11.19). The geomagnetic field is able to produce a correspondingly wide range of induced magnetizations in ordinary crustal rocks. The direction of the induced magnetization is parallel to the Earth’s magnetic field in the rock.

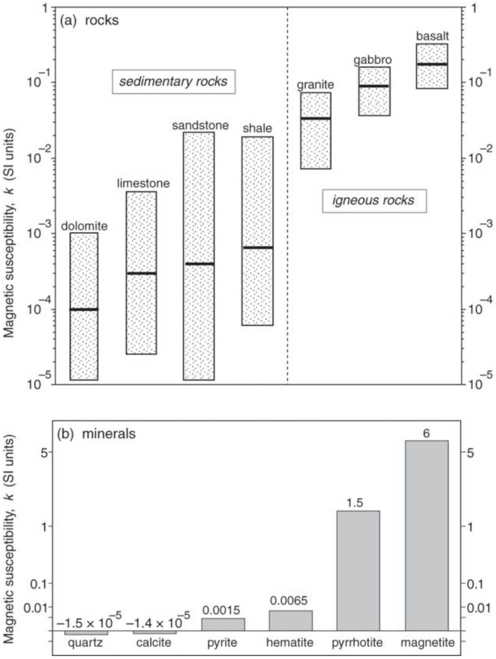


Fig. 11.19 (a) Median values and ranges of the magnetic susceptibility of some common rock types, and (b) the susceptibilities of some important minerals.

The minerals in a rock have different types of magnetic properties, which can be classified as diamagnetic, paramagnetic, or ferromagnetic (Section 12.1 and Appendix C). Each rock usually contains a tiny quantity of ferromagnetic minerals which are more strongly magnetic than the rest of the rock. Grains of these minerals can become magnetized permanently during the formation of the rock or by a later mechanism, acquiring a natural *remanent magnetization* ($\mathbf{M}_r$). This magnetization is not caused by the present-day geomagnetic field, but is related to the Earth's magnetic field in the geological past. Its direction is usually different from that of the present-day field. As a result the directions of $\mathbf{M}_r$ and $\mathbf{M}_i$ are generally not parallel. The direction of $\mathbf{M}_i$ is the same as that of the present field, but the direction of $\mathbf{M}_r$ is often not known unless it can be measured in rock samples.

The total magnetization of a rock is the sum of the remanent and induced magnetizations. As these have different directions they must be combined as vectors (Fig. 11.20a). The direction of the resultant magnetization of the rock is not parallel to the geomagnetic field. If the intensities of $\mathbf{M}_r$ and $\mathbf{M}_i$ are similar, it is difficult to interpret the total magnetization. Fortunately, in many important situations $\mathbf{M}_r$ and $\mathbf{M}_i$ are sufficiently different to permit some simplifying assumptions. The relative importance of the remanent and induced parts of the magnetization is expressed in the *Königsberger ratio* (Q_n), defined as the ratio of the intensity of the remanent magnetization to that of the induced magnetization (i.e., $Q_n = M_r/M_i$).

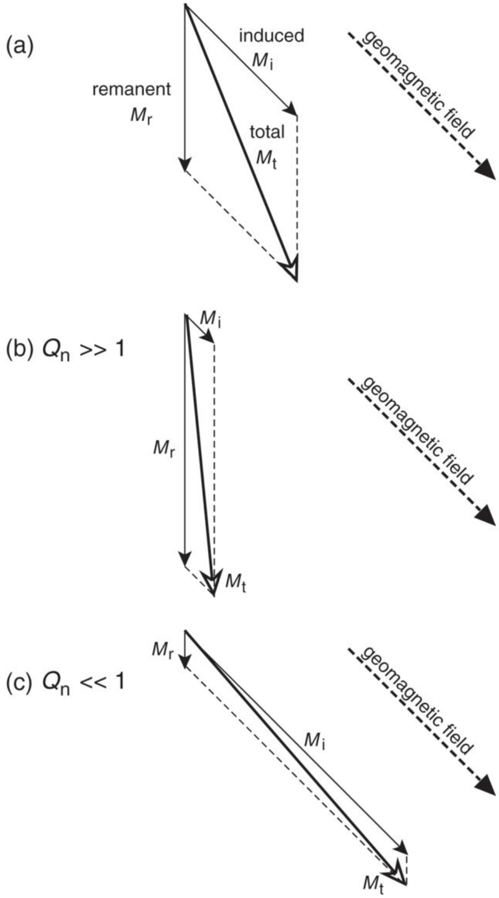

(a)
remanent
M_r
induced
M_i
total
M_t
geomagnetic field
(b) Q_n >> 1
M_i
M_r
M_t
geomagnetic field
(c) Q_n << 1
M_r
M_i
M_t
geomagnetic field

Fig. 11.20 The remanent (M_r), induced (M_i), and total (M_t) magnetizations in a rock. (a) For an arbitrary case Mt lies between M_i and M_r, (b) for a very large Königsberger ratio ($Q_n \gg 1$) M_t is

close to M_r, and (c) for a very small Königsberger ratio ($Q_n \ll 1$) M_t is almost the same as M_i.

Two situations are of particular interest. The first is when Q_n is very large (i.e., $Q_n \gg 1$). In this case (Fig. 11.20b), the total magnetization is dominated by the remanent component and its direction is essentially parallel to $\mathbf{M}_r$. Oceanic basalts, formed by extrusion and rapid underwater cooling at oceanic ridges, are an example of rocks with high Q_n ratios. Due to the rapid quenching of the molten lava titanomagnetite grains form with skeletal structures and very fine grain sizes. The oceanic basalts carry a strong thermoremanent magnetization and often have Q_n values of 100 or greater. This facilitates the interpretation of oceanic magnetic anomalies, because in many cases the induced component can be neglected and the crustal magnetization can be interpreted as if it were entirely remanent.

The other important situation is when Q_n is very small (i.e., $Q_n \ll 1$). This requires the remanent magnetization to be negligible in comparison to the induced magnetization. To minimize its energy the magnetization of a coarse-grained magnetite grain is subdivided into smaller *domains* with antiparallel magnetization directions. The boundaries between the domains are easily moved around by a magnetic field, so the susceptibility is high and the Earth's magnetic field produces a strong induced magnetization. The remanent magnetization is usually weak, because the antiparallel domains tend to cancel each other's magnetization. These two factors yield a low value for Q_n. Magnetic investigations of continental crustal rocks for commercial exploitation (e.g., in ancient shield areas) can often be interpreted as cases with $Q_n \ll 1$. The magnetization can then be assumed to be entirely induced (Fig. 11.20c) and oriented parallel to the direction of the present-day geomagnetic field at the measurement site, which is usually known. This makes it easier to design a model to interpret the feature responsible for the magnetic anomaly.

11.4.2 Magnetometers

The instrument used to measure magnetic fields is called a magnetometer. Until the 1940s magnetometers were mechanical instruments that balanced the torque of the magnetic field on a finely balanced compass needle against a restoring force provided by gravity or by the torsion in a suspension fiber. The balance types were cumbersome, delicate, and slow to operate. For optimum sensitivity they were designed to measure changes in a selected component of the magnetic field, usually the vertical field. This type of magnetometer has now been superseded by more sensitive and robust electronic instruments. The most important of these are the flux-gate, proton-precession, and optically pumped magnetometers.

11.4.2.1 The Flux-Gate Magnetometer

Some special nickel–iron alloys have very high magnetic susceptibility and very low remanent magnetization. Common examples are *Permalloy* (78.5% Ni, 21.5% Fe) and *Mumetal* (77% Ni, 16% Fe, 5% Cu, 2% Cr). The preparation of these alloys involves annealing at very high temperature (1100–1200°C) to remove lattice defects around which internal stress could produce magnetostrictive energy. After this treatment the coercivity of the alloy is very low (i.e., its magnetization can be changed by a very weak field) and its susceptibility is so high that the Earth's field can induce a magnetization in it that is a considerable proportion of the saturation value.

The sensor of a flux-gate magnetometer consists of two parallel strips of the special alloy (Fig. 11.21a). They are wound in opposite directions with primary energizing coils. When a current flows in the primary coils, the parallel strips become magnetized in opposite directions. A secondary coil wound about the primary pair detects the change of magnetic flux in the cores (Fig. 11.21b), which is zero as soon as the cores saturate. While the primary current is rising or falling, the magnetic flux in each strip changes and a voltage is induced in the secondary coil. If there is no external magnetic field, the signals due to the changing flux are equal and opposite and no output signal is recorded. When the axis of the sensor is aligned with the Earth's magnetic field, the latter is added to the primary field in one strip and subtracted from it in the other. The phases of the magnetic flux in the alloy strips are now different; one saturates before the other. The flux changes in the two alloy strips are no longer equal and opposite. An output voltage is produced in the secondary coil that is proportional to the strength of the component of the Earth's magnetic field along the axis of the sensor.

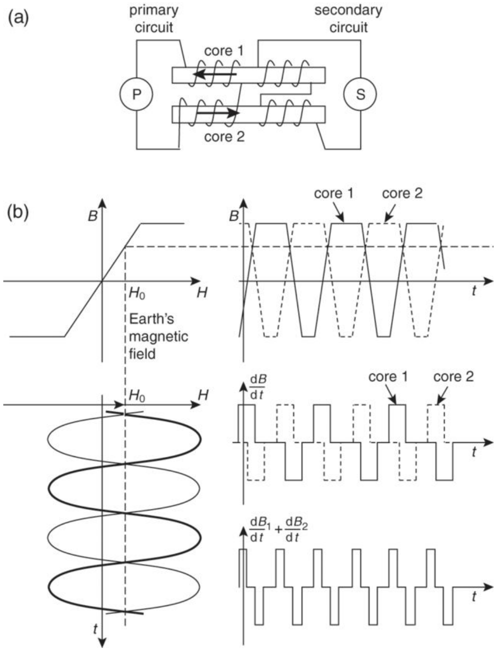

(a)
primary circuit
secondary circuit
core 1
P
S
core 2
(b)
B
H₀
H
Earth's magnetic field
B
core 1
core 2
t
H₀
H
t
dB/dt
core 1
core 2
t
dB₁/dt + dB₂/dt
t

Fig. 11.21 Simplified principle of the flux-gate magnetometer. (a) Primary and secondary electrical circuits include coils wrapped around parallel strips of *Mumetal* in opposite and similar senses, respectively. (b) The output signal in a magnetic field is proportional to the net rate of change of magnetic flux in the *Mumetal* strips.

Source: H. Militzer, R. Scheibe and W. Seiberl, Angewandte Magnetik, in: *Angewandte Geophysik, I: Gravimetrie und Magnetik*,

H. Militzer and F. Weber, eds., pp. 127–189, Vienna: Springer-Verlag, 1984.

The flux-gate magnetometer is a *vector magnetometer*, because it measures the strength of the magnetic field in a particular direction, namely along the axis of the sensor. This requires that the sensor must be accurately oriented along the direction of the field component to be measured. For total field measurements, three sensors are employed. These are fixed at right angles to each other and connected with a feedback system which rotates the entire unit so that two of the sensors detect zero field. The magnetic field to be measured is then aligned with the axis of the third sensor.

The flux-gate magnetometer does not yield absolute field values. The output is a voltage, which must be calibrated in terms of magnetic field. However, the instrument provides a continuous record of field strength. Its sensitivity of about 1 nT makes it capable of measuring most magnetic anomalies of geophysical interest. It is robust and adaptable to being mounted in an airplane, or towed behind it. The instrument was developed during World War II as a submarine detector. After the war it was used extensively in airborne magnetic surveying.

11.4.2.2 The Proton-Precession Magnetometer

Since World War II, sensitive magnetometers have been designed around quantum-mechanical properties. The proton-precession magnetometer depends on the fact that the nucleus of the hydrogen atom, a proton, has a magnetic moment proportional to the angular momentum of its spin. Because the angular momentum is quantized, the proton magnetic moment can only have specified values, which are multiples of a fundamental unit called the nuclear magneton. The situation is analogous to the quantization of magnetic moment associated with electron spin, for which the fundamental unit is the Bohr magneton. The ratio of the magnetic moment to the spin angular momentum is called the *gyromagnetic ratio* (γ_p) of the proton. It is an accurately known fundamental constant with the value $\gamma_p = 2.6752219 \times 10^8\ s^{-1}\ T^{-1}$.

The proton-precession magnetometer is simple and robust in design. The sensor of the instrument consists of a flask containing a proton-rich liquid, such as water. Around the flask are wound a magnetizing solenoid and a detector coil (Fig. 11.22); some designs use the same solenoid alternately for magnetizing and detection. When the current in the magnetizing solenoid is switched on, it creates a magnetic field of the order of 10 mT, which is about 200 times stronger than the Earth's field. The magnetizing field aligns the magnetic moments of the protons along the axis of the solenoid, which is oriented approximately east–west at right angles to the Earth's field. After the magnetizing field is interrupted, the magnetic moments of the proton spins react to the couple exerted on them by the Earth's magnetic field. Like a child's top spinning in the field of gravity, the proton magnetic moments precess about the direction of the ambient magnetic field. They do so at a rate known as the *Larmor precessional frequency*. The motion of the magnetic moments induces a signal in the detector coil. The induced signal is amplified electronically and the precessional frequency is accurately measured by counting cycles for a few seconds. The strength B of the measured magnetic field is directly proportional to the *frequency* of the signal (f), and is given by

$$B = \frac{2\pi}{\gamma_p} f. \tag{11.15}$$

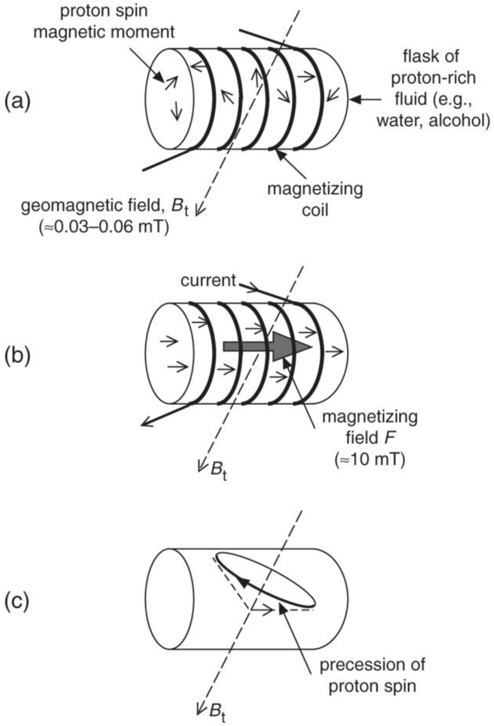

proton spin magnetic moment
(a)
flask of proton-rich fluid (e.g., water, alcohol)
geomagnetic field, B_t (≈0.03–0.06 mT)
magnetizing coil
current
(b)
magnetizing field F (≈10 mT)
B_t
(c)
precession of proton spin
B_t

Fig. 11.22 (a) The elements of a proton-precession magnetometer. (b) Current in the magnetizing coil produces a strong field *F* that aligns the magnetic moments ("spins") of the protons. (c) When the field *F* is switched off, the proton spins precess about the geomagnetic field B_t, inducing an alternating current in the coil with the Larmor precessional frequency *f*.

The intensity of the Earth's magnetic field is in the range 30,000–60,000 nT. The corresponding precessional frequency is approximately 1250–2500 Hz, which is in the audio-frequency range. Accurate measurement of the signal frequency gives an instrumental sensitivity of about 1 nT, but requires a few seconds of observation. Although it gives an absolute value of the field, the proton-precession magnetometer does not give a continuous record.

The flux-gate and proton-precession magnetometers are widely used in magnetic surveying. Their portability and simplicity give them advantages for field use. The two instruments have comparable sensitivities of 0.1–1 nT. In contrast to the flux-gate instrument, which measures the component of the field along its axis, the proton-precession magnetometer cannot measure field components; it is a total-field magnetometer. The total field **B** is the vector sum of the Earth's magnetic field $\mathbf{B}_E$ and the stray magnetic field $\Delta\mathbf{B}$ of, say, an orebody. Generally, $\Delta B \ll B_E$, so that the direction of the total field does not deviate far from the Earth's field. In some applications it is often adequate to regard the measured total field anomaly as the projection of $\Delta\mathbf{B}$ along the Earth's field direction.

11.4.2.3 The Absorption Cell Magnetometer

The absorption cell magnetometer is also referred to as the alkali-vapor or optically pumped magnetometer. The principle of its operation is based on the quantum-mechanical model of the atom. According to their quantum numbers the electrons of an atom occupy concentric shells about the nucleus with different energy levels. The lowest energy level of an electron is its ground state. The magnetic moment associated with the spin of an electron can be either parallel or antiparallel to an external magnetic field. The energy of the electron is different in each case. This results in the ground state splitting into two sublevels with slightly different energies. The energy difference is proportional to the strength of the magnetic field. The splitting of energy levels in the presence of a magnetic field is called the *Zeeman effect.*

Absorption cell magnetometers utilize the Zeeman effect in vapors of alkali elements such as rubidium or cesium, which have only a single valence electron in the outermost energy shell. Consider the schematic representation of an alkali-vapor magnetometer in Fig. 11.23. A polarized light-beam is passed through an absorption cell containing rubidium or cesium vapor and falls on a photoelectric cell, which measures the intensity of the light-beam. In the presence of a magnetic field the ground state of the rubidium or cesium is split into two sublevels, G_1 and G_2. If the exact amount of energy is added to the vapor, the electrons may be raised from their ground state to a higher-energy level, H. Suppose that we irradiate the cell with light from which we have filtered out the spectral line corresponding to the energy needed for the transition G_2H. The energy for the transition G_1H has not been removed, so the electrons in ground state G_1 will receive energy that excites them to level H, whereas those in ground state G_2 will remain in this state. The energy for these transitions comes from the incident light-beam, which is absorbed in the cell. In due course, the excited electrons will fall back to one of the more stable ground states. If an electron in excited state H falls back to sublevel G_1 it will be re-excited into level H; but if it falls back to sublevel G_2 it will remain there. In time, this process – called "optical pumping" – will empty sublevel G_1 and fill level G_2. At this stage no more energy can be absorbed from the polarized light-beam and the absorption cell becomes transparent. If we now supply electromagnetic energy to the system in the form of a radio-frequency signal with just the right amount of energy to permit transitions between the populated G_2 and unpopulated G_1 ground sublevels, the balance will be disturbed. The optical pumping will start up again and will continue until the electrons have been expelled from the G_1 level. During this time energy is absorbed from the light-beam and it ceases to be transparent.

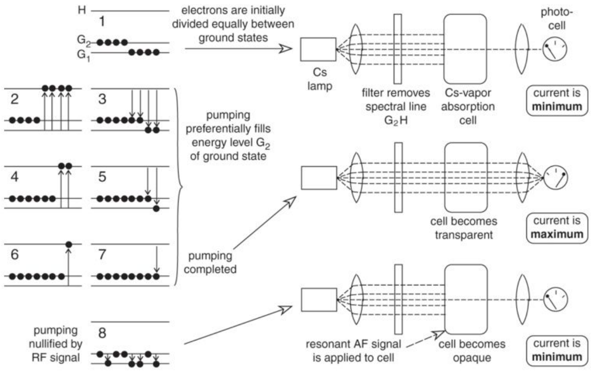

H
1
G2
G1
electrons are initially divided equally between ground states
Cs lamp
filter removes spectral line G2H
Cs-vapor absorption cell
photo-cell
current is minimum
2
3
pumping preferentially fills energy level G2 of ground state
4
5
cell becomes transparent
current is maximum
6
7
pumping completed
pumping nullified by RF signal
8
resonant AF signal is applied to cell
cell becomes opaque
current is minimum

Fig. 11.23 The principle of operation of the optically pumped magnetometer.

Source: W.M. Telford, L.P. Geldart and R.E. Sheriff, *Applied Geophysics*, 770 pp., Cambridge: Cambridge University Press, 1990.

In the rubidium- and cesium-vapor magnetometers, a polarized light-beam is shone at approximately 45° to the magnetic field direction. In the presence of the Earth's magnetic field, the electrons precess about the field direction at the Larmor precessional frequency. At one part of the precessional cycle an electron spin is almost parallel to the field direction, and one half-cycle later it is nearly antiparallel. The varying absorption causes a fluctuation of intensity of the light-beam at the Larmor frequency. This is detected by the photocell and converted to an alternating current. By means of a feedback circuit the signal is supplied to a coil around the container of rubidium gas and a radio-frequency resonant circuit is created. The ambient geomagnetic field B that causes the splitting of the ground state is proportional to the Larmor frequency, and is given by

$$B = \frac{2\pi}{\gamma_e} f . \tag{11.16}$$

Here, γ_e is the gyromagnetic ratio of the electron, which is known with an accuracy of about 1 part in 10^8. It is about 660 times larger than γ_p, the gyromagnetic ratio of the proton. The precessional frequency is correspondingly higher and easier to measure precisely. The sensitivity of an optically pumped magnetometer is very high, about 0.01 nT, which is an order of magnitude more sensitive than the flux-gate or proton-precession magnetometer.

11.4.3 Magnetic Survey Methods

The purpose of magnetic surveying is to identify and describe regions of the Earth's crust that have unusual (anomalous) magnetizations. In the realm of applied geophysics the anomalous magnetizations might be associated with local mineralization that is potentially of commercial interest, or they could be due to subsurface structures that have a bearing on the location of oil deposits. In global geophysics, magnetic surveying over oceanic ridges provided vital clues that led to the theory of plate tectonics and revealed the polarity history of the Earth's magnetic field since the Early Jurassic.

Magnetic surveying consists of: (1) measuring the terrestrial magnetic field at predetermined points; (2) correcting the measurements for known changes; and (3) comparing the resultant value of the field with the expected value at each measurement station. The expected value of the field at any place is taken to be that of the International Geomagnetic Reference Field (IGRF), described in Section 11.2.4. The difference between the observed and expected values is a *magnetic anomaly*.

11.4.3.1 Measurement Methods

The surveying of magnetic anomalies can be carried out on land, at sea, and in the air. In a simple land survey an operator might use a portable magnetometer to measure the field at the surface of the Earth at selected points that form a grid over a suspected geological structure. This method is slow but it yields a detailed pattern of the magnetic field anomaly over the structure, because the measurements are made close to the source of the anomaly.

In practice, the surveying of magnetic anomalies is most efficiently carried out from an aircraft. The magnetometer must be removed as far as possible from the magnetic environment of the aircraft. This may be achieved by mounting the instrument on a fixed boom, A, several meters long (Fig. 11.24a). Alternatively, the device may be towed behind the aircraft in an aerodynamic housing, B, at the end of a cable 30–150 m long. The "bird" containing the magnetometer then flies behind and below the aircraft. The flight environment is comparatively stable. Airborne magnetometers generally have higher sensitivity (≈ 0.01 nT) than those used in ground-based surveying (sensitivity ≈ 1 nT). This compensates for the loss in resolution due to the increased distance between the magnetometer and the source of the anomaly. Airborne magnetic surveying is an economical way to reconnoiter a large territory in a short time. It has become a routine part of the initial phases of the geophysical exploration of an uncharted territory.

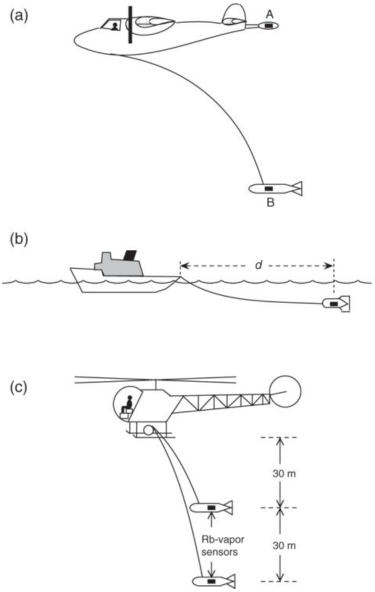


Fig. 11.24 (a) In airborne magnetic surveying the magnetometer may be mounted rigidly on the airplane at the end of a boom (A), or towed in an aerodynamic housing behind the plane (B). (b) In marine studies the magnetometer must be towed some distance *d* behind the ship to escape its magnetic field. (c) A pair of sensitive magnetometers in the same vertical plane act as a magnetic gradiometer (after Slack et al., 1967).

The magnetic field over the oceans may also be surveyed from the air. However, most of the marine magnetic record has been obtained by shipborne surveying. In the marine application a proton-precession magnetometer mounted in a waterproof "fish" is towed behind the ship at the end of a long cable (Fig. 11.24b). Considering that most research vessels consist of several hundred to several thousand tons of steel, the ship causes a large magnetic disturbance. For example, a research ship of about 1000 tons deadweight causes an anomaly of about 10 nT at a distance of 150 m. To minimize the disturbance of the ship the tow-cable must be about 100–300 m in length. At this distance the "fish" in fact "swims" well below the water surface. Its depth is dependent on the length of the tow-cable and the speed of the ship. At a typical survey speed of 10 km h^{-1} its operational depth is about 10–20 m.

11.4.3.2 Magnetic Gradiometers

The magnetic gradiometer consists of a pair of alkali-vapor magnetometers maintained at a fixed distance from each other. In ground-based surveying the instruments are mounted at opposite ends of a rigid vertical bar. In airborne usage two magnetometers are flown at a vertical spacing of about 30 m (Fig. 11.24c). The difference in outputs of the two instruments is recorded. If no anomalous body is present, both magnetometers register the Earth's field equally strongly and the difference in output signals is zero. If a magnetic contrast is present in the subsurface rocks, the magnetometer closest to the structure will detect a stronger signal than the more remote instrument, and there will be a difference in the combined output signals.

The gradiometer emphasizes anomalies from local shallow sources at the expense of large-scale regional variation due to deep-seated sources. Moreover, because the gradiometer registers the *difference* in signals from the individual magnetometers, there is no need to compensate the measurements for diurnal variation, which affect each individual magnetometer equally. Proton-precession magnetometers are most commonly used in ground-based magnetic gradiometers, while optically pumped magnetometers are favored in airborne gradiometers.

11.4.3.3 The Survey Pattern

In a systematic regional airborne (or marine) magnetic survey the measurements are usually made according to a predetermined pattern. In surveys made with fixed-wing aircraft the survey is usually flown at a constant flight elevation above sea level (Fig. 11.25a). This is the procedure favored for regional or national surveys, or for the investigation of areas with dramatic topographic relief. The survey focuses on the depth to the magnetic basement, which often underlies less-magnetic sedimentary surface rocks at considerable depth. In regions that are flat or that do not have dramatic topography, it may be possible to fly a survey at low altitude, as close as possible to the magnetic sources. This method would be suitable over ancient shield areas, where the goal of the survey is to detect local mineralizations with potential commercial value. If a helicopter is being employed, the distance from the magnetic sources may be kept as small as possible by flying at a constant height above the ground surface (Fig. 11.25b).

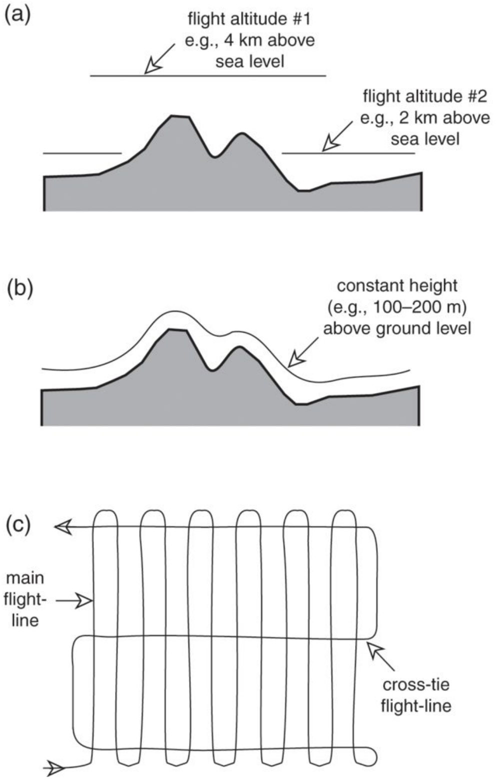

(a)
flight altitude #1
e.g., 4 km above
sea level
flight altitude #2
e.g., 2 km above
sea level
(b)
constant height
(e.g., 100–200 m)
above ground level
(c)
main
flight-
line
cross-tie
flight-line

Fig. 11.25 In airborne magnetic surveying the flight-lines may be flown at (a) constant altitude above sea level, or (b) constant height above ground level. The flight pattern (c) includes parallel measurement lines and orthogonal cross-tie-lines.

The usual method is to survey a region along parallel flight-lines (Fig. 11.25c), which may be spaced anywhere from 100 m to a few kilometers apart, depending on the flight elevation used, the intensity of coverage, and the quality of detail desired. The orientation of the flight-lines is selected to be more or less normal to the trend of suspected or known subsurface features. Additional tie-lines are flown at right angles to the main pattern. Their separation is about 5–6 times that of the main flight-lines. The repeatability of the measurements at the intersections of the tie-lines and the main flight-lines provides a check on the reliability of the survey. If the differences (called closure errors) are large, an area may need to be re-surveyed. Alternatively, the differences may be distributed mathematically among all the observations until the closure errors are minimum.

11.4.4 Reduction of Magnetic Field Measurements

In comparison to the reduction of gravity data, magnetic survey data require very few corrections. One effect that must be compensated for is the variation in intensity of the geomagnetic field at the Earth's surface during the course of a day. This *diurnal variation* is due to the part of the Earth's magnetic field that originates in the ionosphere (Section 11.2.3.2). At any point on the Earth's surface the external field varies during the day as the Earth rotates beneath different parts of the ionosphere. The effect is much greater than the precision with which the field can be measured. The diurnal variation may be corrected by installing a constantly recording magnetometer at a fixed base station within the survey area. Alternatively, the records from a geomagnetic observatory may be used, provided it is not too far from the survey area. The time is noted at which each field measurement is made during the actual survey and the appropriate correction is made from the control record.

The variations of magnetic field with *altitude, latitude*, and *longitude* are dominated by the vertical and horizontal variations of the dipole field. The total intensity B_t of the field is obtained by computing the resultant of the radial component B_r (Eq. 11.10) and the tangential component B_θ (Eq. 11.11):

$$B_t = \sqrt{B_r^2 + B_\theta^2} = \frac{\mu_0 m}{4\pi} \frac{\sqrt{1 + 3\cos^2\theta}}{r^3}. \tag{11.17}$$

The *altitude* correction is given by the vertical gradient of the magnetic field, obtained by differentiating the intensity B_t with respect to radius, *r*. This gives

$$\frac{\partial B_t}{\partial r} = -3 \frac{\mu_0 m}{4\pi} \frac{\sqrt{1 + 3\cos^2\theta}}{r^3} = -\frac{3}{r} B_t. \tag{11.18}$$

The vertical gradient of the field is found by substituting $r = R = 6371$ km and an appropriate value for B_t. It clearly depends on the latitude of the measurement site. At the magnetic equator ($B_t \approx 30{,}000$ nT) the altitude correction is about 0.015 nT m^{-1}; near the magnetic poles ($B_t \approx 60{,}000$ nT) it is about 0.030 nT m^{-1}. The correction is so small that it is often ignored.

In regional studies the corrections for *latitude* and *longitude* are inherent in the reference field that is subtracted. In a survey of a small region, the latitude correction is given by the north–south horizontal gradient of the magnetic field, obtained by differentiating B_t with respect to polar angle, θ. This gives for the northward increase in B_t (i.e., with increasing latitude)

$$-\frac{1}{r}\frac{\partial B_t}{\partial \theta} = \frac{\mu_0 m}{4\pi r^4}\frac{\partial}{\partial \theta}\sqrt{1+3\cos^2\theta}$$

$$= \frac{3B_t \sin\theta \cos\theta}{r(1+3\cos^2\theta)}. \tag{11.19}$$

The latitude correction is zero at the magnetic pole ($\theta = 0°$) and magnetic equator ($\theta = 90°$) and reaches a maximum value of about 5 nT per *kilometer* (0.005 nT m^{-1}) at intermediate latitudes. It is insignificant in small-scale surveys.

In some land-based surveys of highly magnetic terrains (e.g., over lava flows or mineralized intrusions), the disturbing effect of the magnetized topography may be serious enough to require additional topographic corrections.

11.4.5 Magnetic Anomalies

The gravity anomaly of a body is caused by the density contrast ($\Delta\rho$) between the body and its surroundings. The shape of the anomaly is determined by the shape of the body and its depth of burial. Similarly, a magnetic anomaly originates in the magnetization contrast (ΔM) between rocks with different magnetic properties. However, the shape of the anomaly depends not only on the shape and depth of the source object but also on its orientation to the profile and to the inducing magnetic field, which itself varies in intensity and direction with geographical location. In oceanic magnetic surveying the magnetization contrast results from differences in the remanent magnetizations of crustal rocks, for which the Königsberger ratio is much greater than unity (i.e., $Q_n \gg 1$). Commercial geophysical prospecting is carried out largely in continental crustal rocks, for which the Königsberger ratio is much less than unity (i.e., $Q_n \ll 1$) and the magnetization may be assumed to be induced by the present geomagnetic field. The magnetization contrast is then due to susceptibility contrast in the crustal rocks. If k represents the susceptibility of an orebody, k_0 the susceptibility of the host rocks, and H the strength of the magnetic field in the rocks, the magnetization contrast is

$$\Delta M = (k - k_0)H. \tag{11.20}$$

Some insight into the physical processes that give rise to a magnetic anomaly can be obtained from the case of a vertically sided body that is magnetized by a vertical magnetic field. This is a simplified situation because in practice both the body and the field will be inclined, probably at different angles. However, it allows us to make a few observations that are generally applicable. Two scenarios are of particular interest. The first is when the body has a large vertical extent, such that its bottom surface is at a great depth; the other is when the body has a limited vertical extent. In both cases the vertical field magnetizes the body parallel to its vertical sides, but the resulting anomalies have different shapes. To understand the anomaly shapes we will use the concept of magnetic pole distributions.

11.4.5.1 Magnetic Anomaly of a Surface Distribution of Magnetic Poles

Although magnetic poles are a fictive concept, they provide a simple and convenient way to understand the origin of magnetic field anomalies. If a slice is made through a uniformly magnetized object, simple logic tells us that there will be as many south poles per unit of area on one side of the slice as north poles on the opposite side; these will cancel each other and the net sum of poles per unit area of the surface of the slice is zero. This is no longer the case if the magnetization changes across the interface. On each unit area of the surface there will be more poles of the stronger magnetization than poles of the weaker one. A quantitative derivation shows that the resultant number of poles per unit area σ (called the surface density of poles) is proportional to the magnetization contrast ΔM.

Box 11.4 Solid Angles

A solid angle is defined by the ratio between the area of an element of the surface of a sphere and the radius of the sphere. Let the area of a surface element be A and the radius of the sphere be r, as in Fig. B11.4.1. The solid angle Ω subtended by the area A at the center of the sphere is defined as

$$\Omega = \frac{A}{r^2}. \tag{1}$$

The angle subtended by any surface can be determined by projecting the surface onto a sphere. The shape of the area is immaterial. If an element of area *A* is inclined at angle *α* to the radius *r* through a point on the surface, its projection normal to the radius (i.e., onto a sphere passing through the point) is *A*cos*α*, and the solid angle it subtends at the center of the sphere is given by

$$\Omega = \frac{A\cos\alpha}{r^2}. \tag{2}$$

A solid angle is measured in units of steradians, which are analogous to radians in planar geometry. The minimum value of a solid angle is zero, when the surface element is infinitesimally small. The maximum value of a solid angle is when the surface completely surrounds the center of the sphere. The surface area of a sphere of radius *r* is *A* = 4π*r*2 and the solid angle at its center has the maximum possible value, which is 4π.

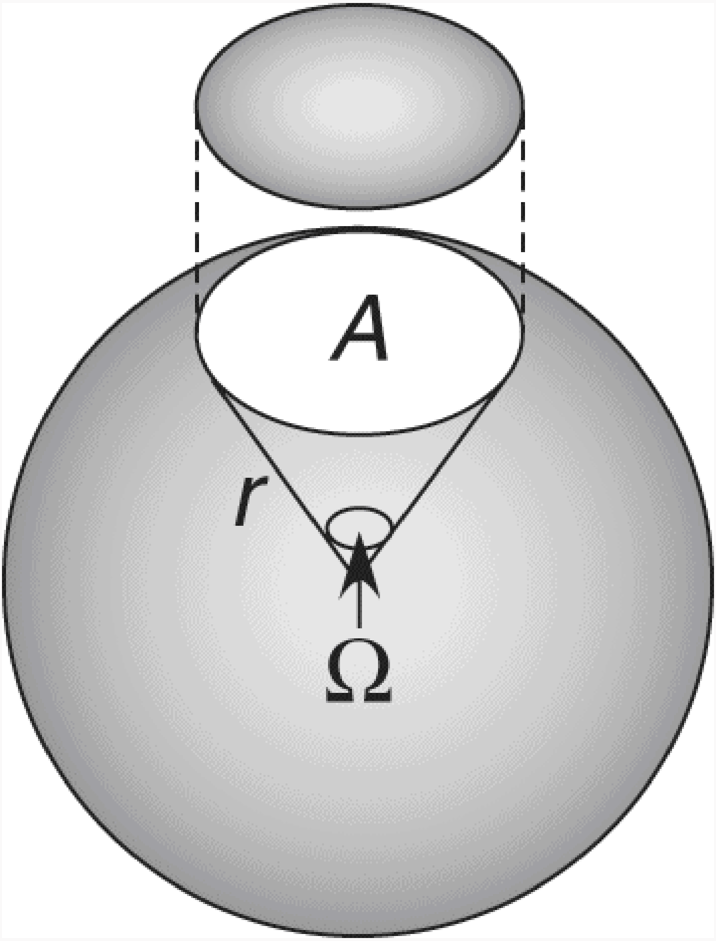


Fig. B11.4.1 Definition of the solid angle Ω subtended by an area *A* on the surface of a sphere with radius *r*.

The concept of the solid angle subtended by a surface element (Box 11.4) provides a qualitative understanding of the magnetic anomaly of a surface distribution of magnetic poles. Consider the distribution of poles on the upper surface with area *A* of a vertical prism with magnetization *M* induced by a vertical field B_z, as illustrated in Fig. 11.26a. At the surface of the Earth, distance *r* from the distribution of poles, the strength of their anomalous magnetic field is proportional to the total number of poles on the surface, which is the product of *A* and the surface density *σ* of poles. The intensity of the field of a magnetic pole decreases as the inverse square of distance *r* (Box 11.1). If the direction of *r* makes an angle θ with the vertical magnetization *M*, the vertical component of the anomalous field at P is found by multiplying by $\cos\theta$. The *vertical magnetic anomaly* ΔB_z of the surface distribution of poles is

$$\Delta B_z \propto \frac{(\sigma A)\cos\theta}{r^2} \propto (\Delta M)\Omega. \tag{11.21}$$

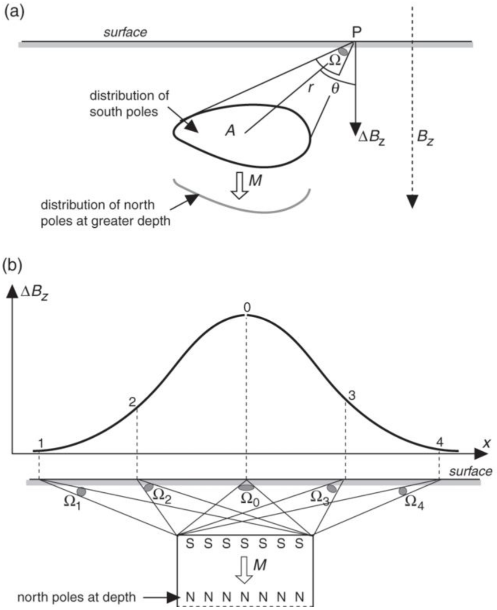

(a)
surface
P
Ω
r
θ
distribution of south poles
A
ΔB_z
B_z
M
distribution of north poles at greater depth
(b)
ΔB_z
0
2
3
1
4
x
surface
Ω_1
Ω_2
Ω_0
Ω_3
Ω_4
S S S S S S S
M
north poles at depth
N N N N N N N

Fig. 11.26 Explanation of the magnetic anomaly of a vertical prism with infinite depth extent. For simplicity the magnetization *M* and inducing field B_z are both assumed to be vertical. (a) The distribution of magnetic poles on the top surface of the prism subtends an angle Ω at the point of measurement. (b) The magnetic anomaly ΔB_z varies along a profile across the prism with the value of the subtended angle Ω.

A more rigorous derivation leads to essentially the same result. At any point on a measurement profile, the magnetic anomaly ΔB_z of a distribution of poles is proportional to the solid angle Ω subtended by the distribution at the point. The solid angle changes progressively along a profile (Fig. 11.26b). At the extreme left and right ends, the radius from the observation point is very oblique to the surface distribution of poles and the subtended angles Ω_1 and Ω_4 are very small; the anomaly distant from the body is nearly zero. Over the center of the distribution, the subtended angle reaches its largest value Ω_0 and the anomaly reaches a maximum. The anomaly falls smoothly on each side of its crest corresponding to the values of the subtended angles Ω_2 and Ω_3 at the intermediate positions. A measurement profile across an equal distribution of "north" poles would be exactly inverted. The north poles create a field of repulsion that acts everywhere to oppose the Earth's magnetic field, so the combined field is less than it would be if the "north" poles were not there. The magnetic anomaly over "north" poles is negative.

11.4.5.2 Magnetic Anomaly of a Vertical Dike

We can now apply these ideas to the magnetic anomaly of a vertical dike. In this and all following examples we will assume a two-dimensional situation, where the horizontal length of the dike (imagined to be into the page) is infinite. This avoids possible complications related to "end effects." Let us first assume that the dike extends to very great depths (Fig. 11.27a), so that we can ignore the effects associated with its remote lower end. The vertical sides of the dike are parallel to the magnetization and no magnetic poles are distributed on these faces. However, the horizontal top face is normal to the magnetization and a distribution of magnetic poles can be imagined on this surface. The direction of magnetization is parallel to the field, so the pole distribution will consist of "south" poles. The magnetized dike behaves like a magnetic *monopole*. At any point above the dike we measure both the inducing field and the *anomalous* "stray field" of the dike, which is directed toward its top. The anomalous field has a component parallel to the Earth's field and so the total magnetic field will be everywhere stronger than if the dike were not present. The magnetic anomaly is everywhere positive, increasing from zero far from the dike to a maximum value directly over it (Fig. 11.27a).

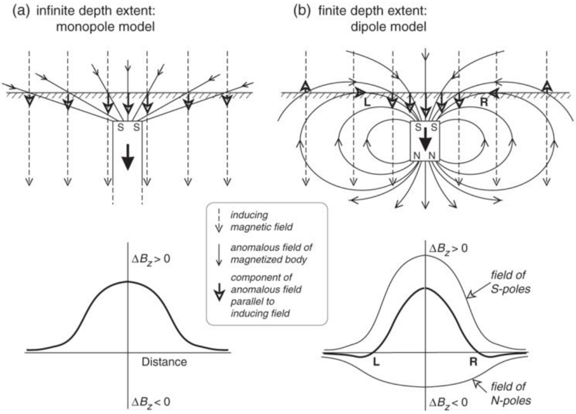

(a) infinite depth extent: monopole model
(b) finite depth extent: dipole model
S S
S S
N N
L
R
inducing magnetic field
anomalous field of magnetized body
component of anomalous field parallel to inducing field
$\Delta B_z > 0$
Distance
$\Delta B_z < 0$
$\Delta B_z > 0$
field of S-poles
L
R
field of N-poles
$\Delta B_z < 0$

Fig. 11.27 (a) The vertical field magnetic anomaly over a vertically magnetized block with infinite depth extent is due only to the distribution of poles on the top surface. (b) If the block has finite depth extent, the pole distributions on the top and bottom surfaces both contribute to the anomaly.

If the vertical extent of the dike is finite, the distribution of north poles on the bottom of the dike may be close enough to the ground surface to produce a measurable stray field. The upper distribution of south poles causes a positive magnetic anomaly, as in the previous example. The lower distribution of north poles causes a negative anomaly (Fig. 11.27b). The north poles are further from the magnetometer than the south poles, so their negative anomaly over the dike is weaker. However, farther along the profile the deeper distribution of poles subtends a larger angle than the upper one does. As a result, the strength of the weaker negative anomaly does not fall off as rapidly along the profile as the positive anomaly does. Beyond a certain lateral distance from the dike (to the left of L and to the right of R in Fig. 11.27b) the negative anomaly of the lower pole distribution is stronger than the positive anomaly of the upper one. This causes the magnetic anomaly to have negative side lobes, which asymptotically approach zero with increasing distance from the dike.

The magnetized dike in this example resembles a bar magnet and can be modeled crudely by a *dipole*. Far from the dike, along a lateral profile, the dipole field lines have a component opposed to the inducing field, which results in the weak negative side lobes of the anomaly. Closer to the dike, the dipole field has a component that reinforces the inducing field, causing a positive central anomaly.

11.4.5.3 Magnetic Anomaly of an Inclined Magnetization

When an infinitely long dike is magnetized obliquely rather than vertically, its anomaly can be modeled either by an inclined dipole or by pole distributions (Fig. 11.28). The magnetization has both horizontal and vertical components, which produce magnetic pole distributions on the vertical sides of the dike as well as on its top and bottom. The symmetry of the anomaly is changed so that the negative lobe of the anomaly is enhanced on the side toward which the horizontal component of magnetization points; the other negative lobe decreases and may disappear.

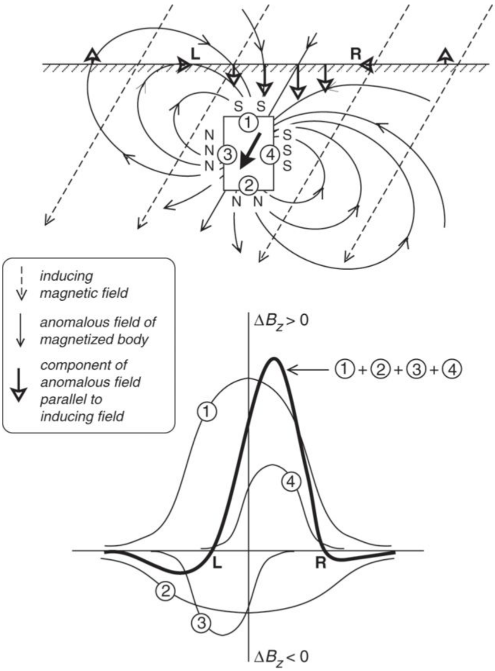

L
R
S S
N N N
S S S
N N
①
②
③
④
inducing magnetic field
anomalous field of magnetized body
component of anomalous field parallel to inducing field
$\Delta B_z > 0$
①+②+③+④
①
④
L
R
②
③
$\Delta B_z < 0$

Fig. 11.28 Explanation of the origin of the magnetic anomaly of an

infinitely long vertical prism in terms of the pole distributions on top, bottom, and side surfaces, when the magnetic field (or magnetization) is inclined.

The shape of a magnetic anomaly also depends on the angle at which the measurement profile crosses the dike, and on the strike and dip of the dike. The geometry, magnetization, and orientation of a body may be taken into account in forward-modeling of an anomaly. However, as in other potential field methods, the inverse problem of determining these factors from the measured anomaly is not unique.

The asymmetry (or *skewness*) of a magnetic anomaly can be compensated by the method of *reduction to the pole*. This consists of recalculating the observed anomaly for the case that the magnetization is vertical. The method involves sophisticated data-processing beyond the scope of this text. The observed anomaly map is first converted to a matrix of values at the intersections of a rectangular grid overlying the map. The Fourier transform of the matrix is then computed and convolved with a filter function to correct for the orientations of the body and its magnetization. The reduction to the pole removes the asymmetry of an anomaly (Fig. 11.29) and allows a better location of the margins of the disturbing body. Among other applications, the procedure has proved to be important for detailed interpretation of the oceanic crustal magnetizations responsible for lineated oceanic magnetic anomalies.

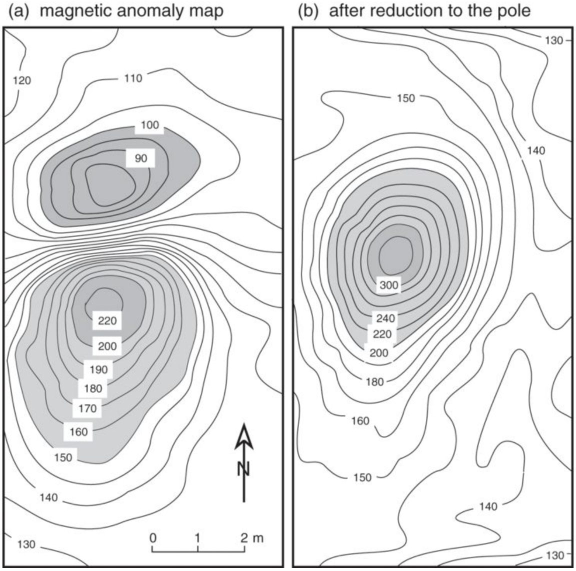

(a) magnetic anomaly map
(b) after reduction to the pole
120
110
100
90
220
200
190
180
170
160
150
140
130
N
0
1
2 m
130
150
140
300
240
220
200
180
160
150
140
130

Fig. 11.29 Effect of data-processing by reduction to the pole on the magnetic anomaly of a small vertical prism with an inclined magnetization. In (a) the contour lines define a dipole-type of anomaly with regions of maximum and minimum intensity (in nT); in (b) the anomaly after reduction to the pole is much simpler and constrains better the location of the center of the prism.

Source: H. Lindner, H. Militzer, R. Rösler and R. Scheibe, Bearbeitung und Interpretation der gravimetrischen und magnetischen Messsergebnisse, in: *Angewandte Geophysik, I: Gravimetrie und Magnetik*, H. Militzer and F. Weber, eds., pp. 127–189, Vienna: Springer-Verlag, 1984

11.4.5.4 Magnetic Anomalies of Simple Geometric Bodies

Box 11.5 Poisson’s Relation

Poisson (1781–1840) observed a relationship between the gravitational and magnetic potentials of a body, which allows a simple way of computing magnetic field anomalies if the gravity anomaly of the body is known. Consider an arbitrary volume *V* with homogeneous density and vertical magnetization (Fig. B11.5). If the density of the body is Δρ, a small element with volume dV has mass (Δρ dV). The gravitational potential U at a point on the surface at a distance r from the element is

$$U = -G\frac{(\Delta\rho)\ \mathrm{d}V}{r}. \tag{1}$$

The vertical gravity anomaly Δg_z of the volume element is found by differentiating U with respect to z:

$$g_z = -\frac{\partial}{\partial z}\left(-G\frac{(\Delta\rho)\mathrm{d}V}{r}\right) = G(\Delta\rho)\ \mathrm{d}V\frac{\partial}{\partial z}\left(\frac{1}{r}\right). \tag{2}$$

If the body is vertically magnetized with uniform magnetization ΔM_z, the magnetic moment of the volume element is ($\Delta M_z dV$). The magnetic moment is directed downward as in Fig. B11.5. The radius vector from the element to a point on the surface makes an angle (π – θ) with the orientation of the magnetization. The magnetic potential W at the point (r, θ) is

$$\begin{aligned} W &= \frac{\mu_0}{4\pi}\frac{(\Delta M_z)\mathrm{d}V\ \cos(\pi-\theta)}{r^2} \\ &= -\frac{\mu_0}{4\pi}\frac{(\Delta M_z)\mathrm{d}V\ \cos\theta}{r^2} = -\frac{\mu_0}{4\pi}\frac{(\Delta M_z)\mathrm{d}V}{r^2}\left(\frac{z}{r}\right). \end{aligned} \tag{3}$$

Note the following relationship:

$$\frac{\partial}{\partial z}\left(\frac{1}{r}\right) = -\frac{1}{r^2}\frac{\partial r}{\partial z} = -\frac{1}{r^2}\frac{\partial}{\partial z}\sqrt{x^2+z^2} = -\frac{1}{r^2}\left(\frac{z}{r}\right). \tag{4}$$

Substituting this result in Eq. (3) gives

$$W = \frac{\mu_0}{4\pi}(\Delta M_z)\mathrm{d}V\frac{\partial}{\partial z}\left(\frac{1}{r}\right). \tag{5}$$

Comparing Eq. (2) and Eq. (5) and eliminating the volume ΔV, we get Poisson's relation:

$$W = \frac{\mu_0}{4\pi}\left(\frac{\Delta M_z}{G\Delta\rho}\right)g_z. \tag{6}$$

This derivation for a small element is also valid for an extended body, as long as the density and magnetization are both uniform.

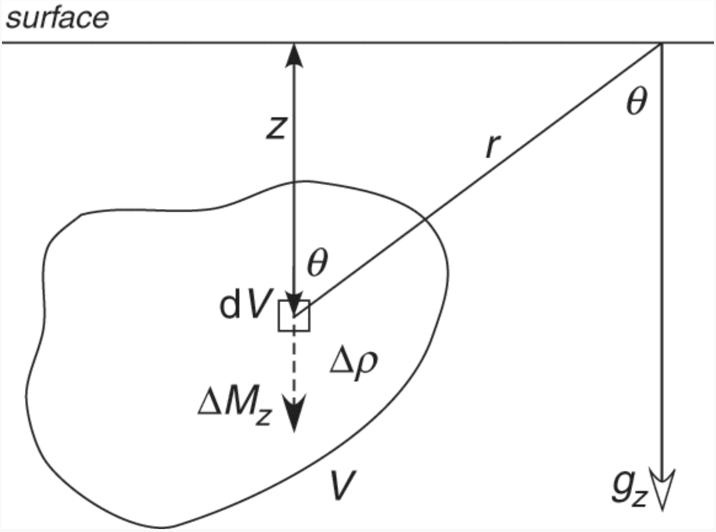


Fig. B11.5 Definition of parameters used in the derivation of Poisson's relation.

The computation of magnetic anomalies is generally more complicated than for gravity anomalies. In practice, iterative numerical procedures are used. However, the Poisson relation (Box 11.5) enables the computation of magnetic anomalies for bodies for which the gravity anomaly is known. This is most easily illustrated for vertically magnetized bodies, such as the following examples.

(1) Sphere

The gravity anomaly Δg_z over a sphere of radius R with density contrast $\Delta\rho$ and center at depth z (representing a diapir or intrusion) is given by Eq. (4.19), repeated here:

$$\Delta g_z = \frac{4}{3}\pi G \Delta\rho R^3 \frac{z}{(z^2+x^2)^{3/2}}. \tag{11.22}$$

Assuming the same dimensions and a magnetization contrast ΔM_z, the potential of the magnetic anomaly over a vertically magnetized sphere according to the Poisson relation is

$$W = \frac{\mu_0}{4\pi}\left(\frac{\Delta M_z}{G\Delta\rho}\right)\Delta g_z = \frac{1}{3}\mu_0 R^3 \Delta M_z \frac{z}{(z^2+x^2)^{3/2}}. \tag{11.23}$$

By differentiating with respect to x or z we get the horizontal or vertical field anomaly, respectively. The vertical field magnetic anomaly ΔB_z of the sphere is

$$\begin{aligned}\Delta B_z &= -\frac{\partial W}{\partial z} = -\frac{1}{3}\mu_0 R^3 \Delta M_z \frac{\partial}{\partial z}\left(\frac{z}{(z^2+x^2)^{3/2}}\right)\\ &= -\frac{1}{3}\mu_0 R^3 \Delta M_z \frac{(z^2+x^2)^{3/2} - z(3/2)(2z)(z^2+x^2)^{1/2}}{(z^2+x^2)^3},\end{aligned} \tag{11.24}$$

$$\Delta B_z = \frac{1}{3}\mu_0 R^3 \Delta M_z \frac{(2z^2 - x^2)}{(z^2 + x^2)^{5/2}}. \tag{11.25}$$

(2) Horizontal Cylinder

The gravity anomaly Δg_z over a cylinder of radius R with horizontal axis centered at depth z and with density contrast $\Delta\rho$ (representing an anticline or syncline) is given by Eq. (4.35). If the structure is vertically magnetized with magnetization contrast ΔM_z, Poisson's relation gives for the magnetic potential

$$W = \frac{\mu_0}{4\pi}\left(\frac{\Delta M_z}{G\Delta\rho}\right)\Delta g_z = \frac{1}{2}\mu_0 R^2 \Delta M_z \frac{z}{(z^2 + x^2)}. \tag{11.26}$$

The vertical magnetic field anomaly ΔB_z over the horizontal cylinder is

$$\Delta B_z = \frac{1}{2}\mu_0 R^2 \Delta M_z \frac{(z^2 - x^2)}{(z^2 + x^2)^2}. \tag{11.27}$$

(3) Horizontal Crustal Block

The gravity anomaly for a thin horizontal sheet of thickness t at depth d between horizontal positions x_1 and x_2 (Fig. 4.23b), extending to infinity normal to the plane of observation, is given by Eq. (4.37). Let the width of the block be $2m$, and let the horizontal position be measured from the mid-point of the block, so that $x_1 = x - m$ and $x_2 = x + m$. Applying Poisson's relation, we get the magnetic potential for a semi-infinite horizontal thin sheet of vertically magnetized dipoles, of thickness t at depth z:

$$W = \frac{\mu_0}{4\pi}\left(\frac{\Delta M_z}{G\Delta\rho}\right)\Delta g_z$$
$$= \frac{\mu_0 t \Delta M_z}{2\pi}\left[\tan^{-1}\left(\frac{x+m}{z}\right) - \tan^{-1}\left(\frac{x-m}{z}\right)\right].$$
(11.28)

Assume that the horizontal crustal block is made up of layers of thickness $t = dz$. If the top of the block is at depth z_1 and its base at depth z_2, the magnetic potential of the block is found by integrating Eq. (11.28) between z_0 and z_1,

$$W = \frac{\mu_0 t \Delta M_z}{2\pi}\int_{z_1}^{z_2}\left[\tan^{-1}\left(\frac{x+m}{z}\right) - \tan^{-1}\left(\frac{x-m}{z}\right)\right]dz.$$
(11.29)

Differentiating with respect to z gives the vertical magnetic field anomaly over the block:

$$\Delta B_z = -\frac{\partial W}{\partial z}$$
$$= -\frac{\mu_0 t \Delta M_z}{2\pi}\int_{z_1}^{z_2}\frac{\partial}{\partial z}\left[\tan^{-1}\left(\frac{x+m}{z}\right) - \tan^{-1}\left(\frac{x-m}{z}\right)\right]dz,$$
(11.30)

$$\Delta B_z = -\frac{\mu_0 t \Delta M_z}{2\pi}\left[\begin{array}{l}\tan^{-1}\left(\frac{x+m}{z_1}\right) - \tan^{-1}\left(\frac{x-m}{z_1}\right)\\ -\tan^{-1}\left(\frac{x+m}{z_2}\right) + \tan^{-1}\left(\frac{x-m}{z_2}\right)\end{array}\right],$$
(11.31)

$$\Delta B_z = \frac{\mu_0 t \Delta M_z}{2\pi}[(\alpha_1 - \alpha_2) - (\alpha_3 - \alpha_4)],$$
(11.32)

where the angles α_1, α_2, α_3, and α_4 are defined in Fig. 11.30a. Note that the angles ($\alpha_1 - \alpha_2$) and ($\alpha_3 - \alpha_4$) are the *planar* angles subtended at the point of measurement by the top and bottom edges of the vertically magnetized crustal block respectively. This is similar to the dependence of magnetic anomalies of three-dimensional bodies on *solid* angles subtended by surfaces of the body at the point of observation (see Fig. 11.26). As was the case for gravity surveys, magnetic profiles normal to the strike of elongate bodies may be regarded as two-dimensional, as long as the third dimension of the body is large enough for variations normal to the profile to be negligible.

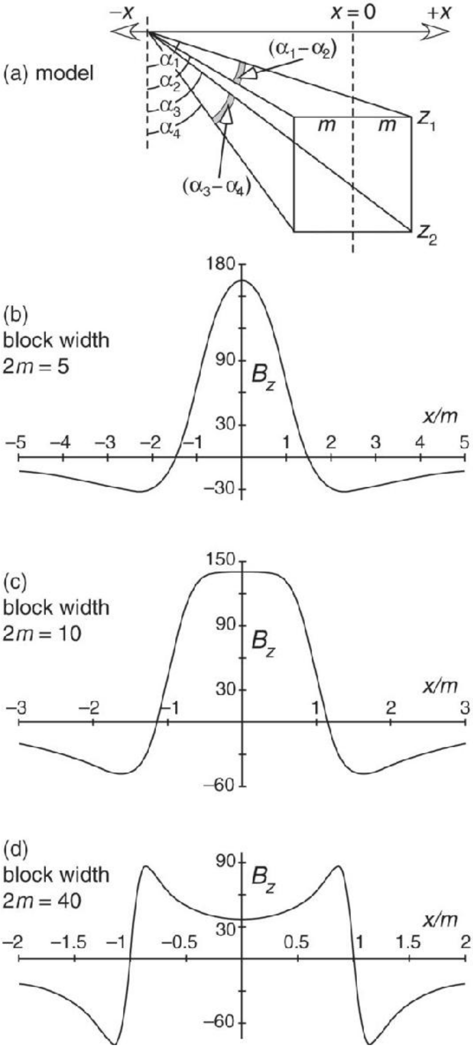


Fig. 11.30 The effect of block width on the shape of the magnetic anomaly over a vertically magnetized thin crustal block. (a) The block has width $w = 2m$, and its top and bottom surfaces are at depths z_1 and z_2, respectively. For each of the calculated anomalies in (b), (c), and (d) these depths are $z_1 = 2.5$ km and $z_2 = 3$ km, and only the width of the block is varied. The amplitude of the anomaly is in arbitrary units.

11.4.5.5 Effect of Block Width on Anomaly Shape

The effect of the width of a crustal block on anomaly shape is illustrated by using Eq. (11.31) to model the vertical field magnetic anomaly of a vertically magnetized block with its top at depth 2.5 km and base at depth 3 km. The block is effectively a thin magnetized layer, similar to the source of oceanic magnetic anomalies. Three cases are considered here: a narrow block of width $w = (2m) = 5$ km for which $m/z_1 = 1$; a block of width 10 km ($m/z_1 = 2$); and a wide block of width 40 km ($m/z_1 = 8$).

The narrowest block gives a sharp, positive central anomaly with negative side lobes (Fig. 11.30b), as explained in Section 11.4.5.2. As the block widens with respect to its depth, the top of the central anomaly flattens (Fig. 11.30c), its amplitude over the middle of the block decreases, and the negative side lobes grow. When the block is much wider than the depth to its top (Fig. 11.30d), a dip develops over the center of the block. The positive anomalies are steep-sided and are maximum just within the edges of the block, while the null values occur close to the edges of the block. The negative side anomalies are almost as large as the positive anomalies.

The pronounced central dip in the anomaly is due to the limited vertical thickness of the layer. If the layer is very wide relative to its thickness, the central anomaly may diminish almost to zero. This is because the angle ($\alpha_3 - \alpha_4$) subtended by the magnetized base of the layer is almost (but not quite) as large as the angle ($\alpha_1 - \alpha_2$) subtended by the top of the layer. For a very large width-to-thickness ratio, the central anomaly is zero, the edge anomalies separate and become equivalent to separate anomalies over the edges of the block.

Examination of Fig. 11.30a shows that the subtended angles ($\alpha_1 - \alpha_2$) and ($\alpha_3 - \alpha_4$), and thus the anomaly shape, depend also on the height of the measurement profile above the surface of the block. A low-altitude profile over the block will show a large central dip, while a high-altitude profile over the same block will show a smaller dip or none at all.

In contrast to the example of a thin layer described above, if the crustal block is very thick, extending to great depth, the angle ($\alpha_3 - \alpha_4$) is zero and the effects of the magnetization discontinuity (or pole distribution) on its base are absent. The shape of the anomaly is then determined by ($\alpha_1 - \alpha_2$) and is flat-topped over a wide block.

11.4.6 Oceanic Magnetic Anomalies

In the late 1950s marine geophysicists conducting magnetic surveys of the Pacific Ocean basin off the west coast of North America discovered that large areas of oceanic crust are characterized by long stripes of alternating positive and negative magnetic anomalies. The striped pattern is best known from studies carried out across oceanic ridge systems (see Fig. 2.10). The striped anomalies are hundreds of kilometers in length parallel to the ridge axis, 10–50 km in width, and their amplitudes amount to several hundreds of nanotesla. On magnetic profiles perpendicular to a ridge axis the anomaly pattern is found to exhibit a remarkable symmetry about the axis of the ridge. The origin of the symmetric lineated anomaly pattern cannot be explained by conventional methods of interpretation based on susceptibility contrast.

Seismic studies indicate a layered structure for the oceanic crust. The floor of the ocean lies at water depths of 2–5 km, and is underlain by a layer of sediment of variable thickness, called seismic Layer 1. Under the sediments lie a complex of basaltic extrusions and shallow intrusions, about 0.5 km thick, forming seismic Layer 2A, under which are found the deeper layers of the oceanic crust consisting of a complex of sheeted dikes (Layer 2B) and gabbro (Layer 3). The magnetic properties of these rocks were first obtained by studying samples dredged from exposed crests and ridges of submarine topography. The rocks of Layers 2B and 3 are much less magnetic than those of Layer 2A. Samples of pillow basalt dredged near to oceanic ridges have been found to have moderate susceptibilities for igneous rocks, but their remanent magnetizations are intense. Their Königsberger ratios are commonly in the range 5–50 and frequently exceed 100. Recognition of these properties provided the key to understanding the origin of the lineated magnetic anomalies. In 1963 the English geophysicists F. J. Vine and D. H. Matthews proposed that the *remanent magnetizations* (and not the susceptibility contrast) of oceanic basaltic Layer 2 were responsible for the striking lineated anomaly pattern. This hypothesis soon became integrated into a working model for understanding the mechanism of sea-floor spreading (see Section 2.5 and Fig. 2.10).

The oceanic crust formed at a spreading ridge acquires a thermoremanent magnetization (TRM) in the geomagnetic field. This is an important form of permanent magnetization in rocks (described in Section 12.1.7.1). The basalts in Layer 2 A are sufficiently strongly magnetized to account for most of the anomaly measured at the ocean surface. For a lengthy period of time (measuring several tens of thousands to millions of years) the polarity of the field remains constant; crust formed during this time carries the same polarity as the field. After a polarity reversal, freshly formed basalts acquire a TRM parallel to the new field direction, i.e., opposite to the previous TRM. Adjacent oceanic crustal blocks of different widths, determined by the variable time between reversals, carry antiparallel remanent magnetizations.

The oceanic crust is magnetized in long blocks parallel to the spreading axis, so the anomaly calculated for a profile perpendicular to the axis is two-dimensional, as in the previous examples. Consider the case where the anomalies on a profile have been reduced to the pole, so that their magnetizations can be taken to be vertical. We can apply the concept of magnetic pole distributions to each block individually to determine the shape of its magnetic anomaly (Fig. 11.31a). If the blocks are contiguous, as is the case when they form by a continuous process such as sea-floor spreading, their individual anomalies will overlap (Fig. 11.31b). The spreading process is symmetric with respect to the ridge axis, so a mirror-image of the sequence of polarized blocks is formed on the other side of the axis (Fig. 11.31c). If the two sets of crustal blocks are brought together at the spreading axis, a magnetic anomaly sequence ensues that exhibits a symmetric pattern with respect to the ridge axis (Fig. 11.31d).

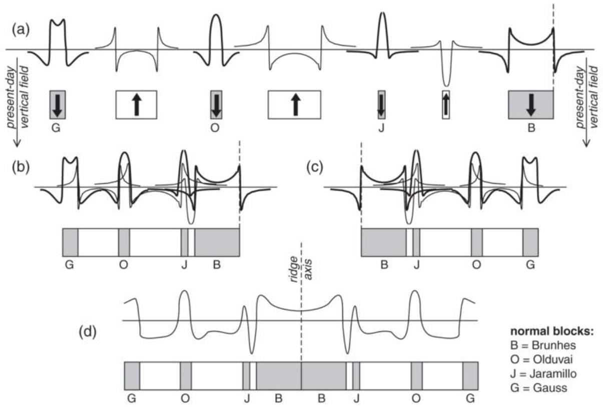


Fig. 11.31 Explanation of the shape of a magnetic profile across an oceanic spreading center: (a) the anomalies of individual oppositely magnetized crustal blocks on one side of the ridge; (b) overlap of the individual anomalies; (c) the effect for the opposite sequence of blocks on the other side of the ridge; and (d) the complete anomaly profile.

This description of the origin of oceanic magnetic anomalies is over-simplified, because the crustal magnetization is more complicated than assumed in the block model. For example, the direction of the remanent magnetization, acquired at the time of formation of the ocean crust, is generally not the same as the direction of the magnetization induced by the present-day field. However, the induced magnetization has uniformly the same direction in the magnetized layer, which thus behaves like a uniformly magnetized thin horizontal sheet and does not contribute to the magnetic anomaly. Moreover, oceanic rocks have high Königsberger ratios, and so the induced magnetization component is usually negligible in comparison to the remanent magnetization. An exception is when a magnetic survey is made close to the magnetized basalt layer, in which case a topographic correction may be needed.

Unless the strike of a ridge is north–south, the magnetization inclination must be taken into account. Skewness is corrected by reducing the magnetic anomaly profile to the pole (Section 11.4.5.3). A possible complication may arise if the oceanic magnetic anomalies have two sources. The strongest anomaly source is doubtless basaltic Layer 2B, but, at least in some cases, an appreciable part of the anomaly may arise in the deeper gabbroic Layer 3. The two contributions are slightly out of phase spatially, because of the curved depth profiles of cooling isotherms in the oceanic crust. This causes a magnetized block in the deeper gabbroic layer to lie slightly further from the ridge than the corresponding block with the same polarity in the basaltic layer above it. The net effect is an asymmetry of inclined magnetization directions on opposite sides of a ridge, so that the magnetic anomalies over blocks of the same age have different skewnesses.

11.5 Suggestions for Further Reading

Introductory Level

Campbell, W. H. 2001. *Earth Magnetism: A Guided Tour Through Magnetic Fields*. San Diego, CA: Harcourt/Academic Press.

Campbell, W. H. 2003. *Introduction to Geomagnetic Fields*. Cambridge: Cambridge University Press.

Kearey, P., Brooks, M., and Hill, I. 2002. *An Introduction to Geophysical Exploration* (3rd ed.). Oxford: Blackwell Publishing.

Intermediate and Advanced Level

Blakely, R. J. 1995. *Potential Theory in Gravity and Magnetic Applications*. Cambridge: Cambridge University Press.

Gubbins, D. and Herrero-Bervera, E. (eds.) 2007. *Encyclopedia of Geomagnetism and Paleomagnetism*. Dordrecht: Springer.

11.6 Review Questions

1. What is the evidence that the geomagnetic field originates inside the Earth?

2. Why is the geomagnetic field at the Earth's surface mainly a dipole field?

3. What is the non-dipole geomagnetic field? How large is it compared to the dipole field at the Earth's surface? Is its importance relative to the dipole field greater or less at the core–mantle boundary?

4. What is the magnetosphere? How does it originate?

5. What are the Van Allen belts? How are they formed?

6. Why do electrically charged particles from the solar wind have curved paths in the Earth's magnetic field?

7. The geomagnetic field has a large anomaly over the South Atlantic, where its intensity is weakened by about 20%. What effects could this anomaly have on extra-terrestrial radiation that impinges on the Earth? What might be the consequences for (a) Earth-orbiting satellites, (b) astronauts, (c) passengers and crew in high-flying aircraft?

8. Which planets do not have a global magnetic field?

9. What is the ionosphere? What effect does it have on the magnetic field measured at the Earth's surface?

10. What is a *magnetic storm*? What is its cause and what are its effects?

11. Describe the principle of operation of (a) the flux-gate magnetometer, and (b) the proton-precession magnetometer.

12. Which corrections must be made to magnetic survey measurements in order to define a magnetic anomaly?

13. Describe how the magnetic anomaly of a very wide, vertically magnetized thin sheet varies with position from one side of the sheet to the other. Where is the anomaly largest and where is it smallest?

11.7 Exercises

1. The IGRF for 2015 gives the following Gauss coefficients for the geomagnetic dipole: g_1^0 = –29,442 nT, g_1^1 = –1501 nT, h_1^1 = + 4797 nT.

(a) Show that the Earth's dipole magnetic moment is 7.724×10^{22} A m^2.

(b) Calculate the locations of the geomagnetic poles.

2. Assuming an *axial* dipole with magnetic moment 7.724×10^{22} A m^2, calculate the variation of the total field with latitude in nT/ km at 30°N.

3. Compute the inclination and declination of the magnetic field that would be observed in Boulder, Colorado (40°N, 105°W) if the Earth's field corresponded to a perfect geocentric dipole whose axis penetrates the Earth's surface at 80°N, 72°W.

4. Show that, for a small displacement along a meridian of magnetic longitude at magnetic latitude 45°N, the change of inclination is exactly four-fifths of the change in latitude.

5. The north magnetic dip pole is at 86.3°N, 160.1°W and the south magnetic dip pole is at 64.3°S, 136.6°E.

(a) Why are the poles not antipodal (exactly opposite)?

(b) What is the closest distance (in km) between the center of the Earth and the straight line joining the poles?

6. Measurements of the magnetic field elements at a geomagnetic observatory gave the following results:

N-component = 27,200 nT; E-component = –1800 nT; V-component = –40,100 nT.

(a) Is the observatory in the northern or southern hemisphere?

(b) What is the total field intensity at the site?

(c) What are the local values of inclination and declination of the field?

7. The IGRF for 2015 gives the values in nT of the Gauss coefficients for the dipole and quadrupole components of the geomagnetic field shown in the following table.

g_1^0	−29,442	g_2^0	−2445
g_1^1	−1501	g_2^1	+3013
h_1^1	+4797	h_2^1	−2846
		g_2^2	+1677
		h_2^2	−642

The mean squared value, R_n, of the intensity of the geomagnetic field at Earth's surface due to the component of degree n is given by

$$R_n = (n+1)\sum_{m=0}^{n}\left((g_n^m)^2 + (h_n^m)^2\right).$$

Using the values in the table:

(a) Calculate the root mean square intensity of the dipole field at the Earth's surface.

(b) Calculate the corresponding intensity of the quadrupole component and express it as a percentage of the dipole field intensity.

8. The values of the Gauss coefficients in the previous exercise are given for the Earth's surface. Re-calculate the root mean square intensities of the dipole and quadrupole field components at the core–mantle boundary (Earth's radius = 6371 km, core radius = 3485 km). How does their ratio compare with that at the Earth's surface?

9. What are the values of n and m in the designation Y_n^m for the spherical harmonic functions illustrated in Fig. B11.3.1 in Box 11.3? Sketch how these patterns would appear on the opposite side of the reference sphere to the one you are looking at?

10. In an aeromagnetic survey at a flight altitude of 2000 m above sea level, the maximum total field anomaly over an orebody was 30 nT. In a repeated measurement at 2500 m altitude the maximum amplitude of the anomaly was 20 nT. Calculate the depth of the orebody below sea level assuming (a) a monopole source and (b) a dipole source for the anomaly.

11. The vertical field magnetic anomaly ΔB_z over a vertically magnetized anticline (represented by a horizontal cylinder) is given by Eq. (11.27). Draw a sketch of the anomaly on a profile normal to the structure. Observe the horizontal positions where the anomaly is zero.

(a) Calculate the horizontal positions where the anomaly has extreme values.

(b) Calculate the peak-to-peak values of the anomaly.

12. Assume that the core of an anticline consists of basalt with susceptibility 1.8×10^{-1} (SI), the host formation is limestone with susceptibility 3×10^{-4} (SI), and the rocks are vertically magnetized.

(a) Compute the induced magnetization contrast when the vertical magnetic field intensity is 40,000 nT.

(b) If the anticline is modeled by a horizontal cylinder whose radius is one-fifth the depth of its axis, calculate the maximum amplitude of the vertical field anomaly over the structure.

13.

(a) Assume the gravity anomaly over an anticline (Eq. 4.35) and apply the Poisson relationship (Box 11.5) to obtain the *horizontal* field magnetic anomaly ΔB_x of a vertically magnetized anticline with radius R and magnetization contrast ΔM_z.

(b) Sketch the shape of the anomaly and determine the horizontal positions of its extreme values.

11.8 Computational Exercises

Jupyter notebook computational exercises can be found online for this chapter at www.cambridge.org/FoG3ed (Magnetic anomalies).

12

Paleomagnetism

Preview

Rocks contain a tiny proportion of magnetic minerals that make them weakly magnetic. Some are magnetized in the direction of the magnetic field in which they formed. By analyzing this direction, the position of a virtual geomagnetic pole (VGP) at the time the rock formed can be located. Connecting the VGP for rocks with different ages from the same continent gives a curve of apparent polar wander (APW), which results from the motion of the continent relative to the rotation axis. Comparing APW paths for different continents reveals a history of relative motions and allows the reconstruction of past supercontinents. The paleomagnetic field has reversed polarity numerous times in the geological past, leaving a record of geomagnetic polarity in the magnetizations of rocks. Polarity reversals cause oceanic magnetic anomalies, which are important for understanding plate tectonics and also form the basis of a geomagnetic polarity timescale for the past 230 Myr.

12.1 Rock Magnetism

12.1.1 Introduction

Paleomagnetism can determine quantitatively the relative locations, history of accretion and fragmentation, and the relative movements of tectonic plates in the geological past. It can define a magnetic polarity stratigraphy that correlates geological events and helps to refine the geological timescale. These contributions are possible because certain rocks become magnetized by the geomagnetic field at the time when they are formed, and the magnetization they acquire is very stable, remaining unchanged for millions of years. The magnetization is carried by a small number of minerals that behave like tiny permanent magnets. As a result, our ability to decipher the geological history of plate motions and continental displacements depends on the magnetic properties of rocks. We must therefore understand their magnetizations and how they are acquired.

Problems in magnetism are often more complicated than those in gravitation and electrostatics. Gravitational and electrostatic fields are central forces, which vary as the inverse square of distance from a center of attraction. Magnetic fields do not act radially toward a central point. Moreover, even in the simplest case (that of a magnetic dipole or a small current loop) the field strength falls off inversely as the *cube* of distance. To complicate matters further, we have to take account of *two* magnetic fields, **B** and **H**.

All magnetic fields originate with electrical currents. This is the case even for permanent magnets, in which currents are associated with the motions of electrons about atomic nuclei. The fundamental magnetic field associated with currents in any medium is **B**, which is called the magnetic induction or magnetic flux density. It is defined by Lorentz’s law and is measured in units of tesla with dimensions N A^{-1} m^{-1} (Box 11.1). The quantity **H** is a computational parameter, which describes how **B** is modified by the magnetization inside a magnetizable material.

12.1.1.1 Magnetization and the Magnetic Field Inside a Material

The magnetic moment **m** of each atom in a material is a vector associated with a current loop (Section 11.1). The net magnetic moment of a volume *V* of the material depends on the degree of alignment of the individual atomic magnetic moments. It is the vector sum of all the atomic magnetic moments in the material. The magnetic moment per unit volume of the material is called its *magnetization*, denoted **M**:

$$\mathbf{M} = \frac{1}{V}\sum_k \mathbf{m}_k. \tag{12.1}$$

Magnetization has the dimensions of magnetic moment (A m^2) divided by volume (m^3), so the SI units of **M** are A m^{-1}. The dimensions of **B** are N A^{-1} m^{-1} and those of μ_0 are N A^{-2} (Box 11.1); consequently the dimensions of B/μ_0 are also A m^{-1}. In general, the magnetization **M** inside a magnetic material will not be exactly equal to B/μ_0; let the difference be **H**, then

$$\mathbf{H} = \mathbf{B}/\mu_0 - \mathbf{M}. \tag{12.2}$$

Magnetization is a property of materials. In a vacuum the magnetization is zero ($\mathbf{M} = 0$), the vectors **B** and **H** are then parallel and proportional ($\mathbf{B} = \mu_0 \mathbf{H}$). In a material **M** is parallel and proportional to **H**:

$$\mathbf{M} = k\mathbf{H}. \tag{12.3}$$

The proportionality factor *k* is a physical property of the material, called the *magnetic susceptibility*. It is a measure of the ease with which the material can be magnetized. Because **M** and **H** have the same dimensions (A m^{-1}), k is a dimensionless quantity. In general, Eq. (12.2) can be written as

$$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M}) = \mu_0\mathbf{H}(1 + k), \tag{12.4}$$

$$\mathbf{B} = \mu\mu_0\mathbf{H}. \tag{12.5}$$

The quantity $\mu = (1 + k)$ is called the *magnetic permeability* of the material. The term "permeability" recalls the early nineteenth-century association of magnetic powers with an invisible fluid. The permeability of a material expresses the ability of the material to allow a fluid to pass through it. Likewise, magnetic permeability is a measure of the ability of a material to convey a magnetic flux. Ferromagnetic metals have high permeabilities; in contrast, minerals and rocks have low susceptibilities and permeabilities $\mu \approx 1$.

The difference between the **B**-field and the **H**-field is evident from the configurations of their respective field lines. The field lines of **B** always form continuous loops, while the field lines of **H** are discontinuous at surfaces where the magnetization **M** changes in strength or direction. This is illustrated for a uniformly magnetized sphere in Fig. 12.1. The **H**-field begins and ends at surfaces where the magnetization **M** is interrupted; inside the sphere **H** acts in the *opposite* direction to the **B**-field. Outside the magnet the **B**-field and **H**-field are parallel; thus the **H**-field is discontinuous at the surface of the magnet.

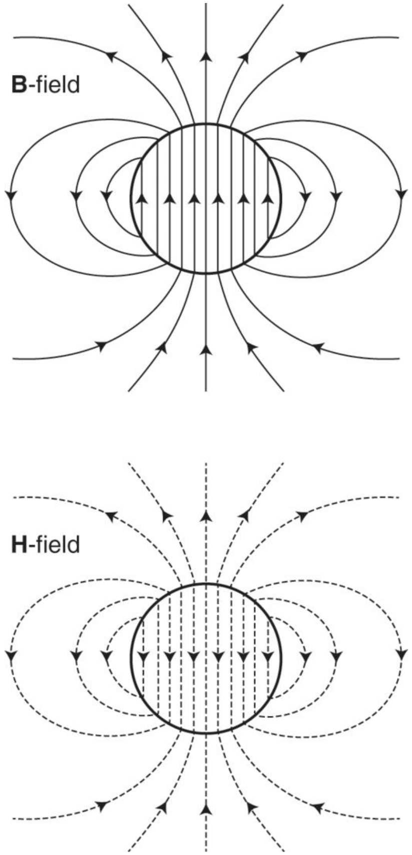


Fig. 12.1 Field lines of the magnetic **B**-field and **H**-field inside and outside a uniformly magnetized sphere. The fields outside the sphere correspond to a dipole.

12.1.2 Mineral Magnetism

The magnetic behavior of a solid is determined by the magnetic moments of the atoms or ions it contains. An individual magnetic moment is proportional to the quantized angular momenta associated with the orbital motion of electrons about the nucleus and with the spins of the electrons about their own axes of rotation. The atomic orbits are arranged around the nucleus in shells, in which the magnetic moments of paired opposite spins cancel each other out. Consequently, the net magnetic moment of a filled shell is zero; the magnetic moment of an atom or ion arises from incompletely filled shells that contain unpaired spins. Inside a solid the atoms or ions are not randomly distributed but occupy fixed positions in a regular lattice, which reflects the solid's internal symmetry and controls interactions between the ions. Hence, the different types of magnetic behavior observed in solids depend not only on the presence of ions with unpaired spins, but also on the lattice symmetry and cell size.

Three main classes of magnetic behavior can be defined on the basis of their magnetic susceptibility. These are diamagnetism, paramagnetism, and ferromagnetism (Appendix C). *Diamagnetism* is characterized by a susceptibility that is weak and negative, i.e., a magnetization develops in the opposite direction to an applied field. All materials show a diamagnetic reaction in a magnetic field, but the diamagnetism is often masked by stronger paramagnetic or ferromagnetic properties. It is observed in materials in which the electron spins are paired; e.g., every spin is canceled by an oppositely oriented spin.

Paramagnetism develops in materials in which one or more electron spins in each atom is unpaired, so that the net magnetic moment of an atom or ion is no longer zero. The resultant magnetic moment can align with a magnetic field. However, the individual atoms react independently; there is no coordinated response to the applied field and so paramagnetic materials have weak, positive susceptibilities.

Ferromagnetism is characterized by magnetic hysteresis (Appendix C) and high susceptibility. It is observed in metals like iron, nickel, and cobalt, in which the lattice geometry and spacing allows the direct exchange of electrons between neighboring atoms. This gives rise to a *molecular field* which aligns the magnetic moments of adjacent atoms parallel to a common direction.

Similar magnetic behavior is observed in a small number of crystalline materials, where it is called *ferrimagnetism*. In contrast to a metal, the ions in a mineral are kept apart by the oxygen ions that form the crystal structure. This makes a direct exchange of electrons impossible, However, in a few minerals the crystal structure permits an *indirect exchange*. This confers magnetic properties on the mineral that are similar to those of ferromagnetism.

Ferrimagnetic minerals are important in geophysics because they have high susceptibilities and give rise to magnetic anomalies induced by the present-day geomagnetic field. They are also able to acquire a remanent magnetization and record the Earth’s magnetic field in the geological past. The most important ferrimagnetic rock-forming mineral is magnetite. It has a strong spontaneous magnetization, a Curie temperature, and the ability to acquire a remanent magnetization (Appendix C). Another important magnetic mineral is hematite. However, its magnetization may be acquired at a later time than when the rock was formed, and this timing uncertainty can restrict its usefulness for studying the magnetic field in the geological past. Pyrrhotite, maghemite, greigite, and goethite can behave like ferrites, but – unlike hematite – they are unstable, converting to other mineral forms.

A rock is a heterogeneous assemblage of different minerals. The matrix minerals are mainly silicates or carbonates, which are diamagnetic in character. Interspersed in this matrix is a lesser quantity of secondary minerals (such as the clay minerals) that have paramagnetic properties. Most of the minerals in a rock contribute to the magnetic susceptibility but are incapable of any contribution to the remanent magnetization, which is due to a dilute dispersion of ferrimagnetic minerals (e.g., commonly less than 0.01% in a limestone). The variable concentrations of ferrimagnetic and matrix minerals result in a wide range of susceptibilities in rocks (Fig. 11.19).

The most important magnetic minerals that occur in rocks are iron–titanium oxides, which are naturally occurring ferrites. They are classified on the basis of their composition and mineral structure as a ternary system, which is a triangular plot that shows graphically the relative compositions of minerals that belong to the system in terms of three components.

12.1.3 The Ternary Oxide System of Magnetic Minerals

The mineral structure of iron–titanium oxides consists of a close-packed lattice of oxygen ions, in which some of the interstitial spaces are occupied by regular arrays of ferrous (Fe^{2+}) and ferric (Fe^{3+}) iron ions and titanium (Ti^{4+}) ions. The relative proportions of these three ions determine the ferrimagnetic properties of the mineral. The composition of an iron–titanium oxide can be illustrated graphically on the ternary oxide diagram (Fig. 12.2), the corners of which represent the minerals rutile (TiO_2), wustite (FeO), and hematite (Fe_2O_3). The proportions of these three constituents in an iron oxide mineral define a point on the ternary diagram. The vertical distance of the point above the FeO–Fe_2O_3 baseline reflects the amount of titanium in the lattice. Hematite is in a higher state of oxidation than wustite, thus the horizontal position along the FeO–Fe_2O_3 axis expresses the degree of oxidation.

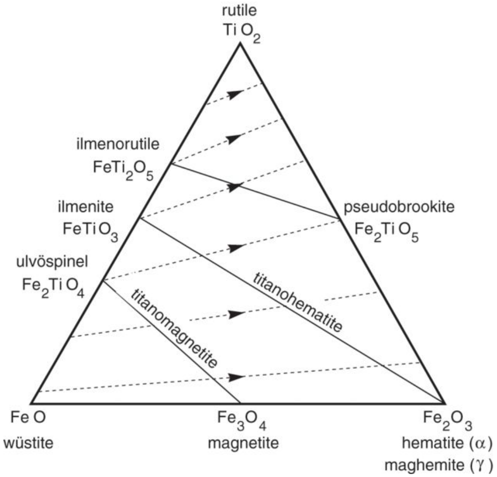

rutile
TiO_2
ilmenorutile
$FeTi_2O_5$
ilmenite
$FeTiO_3$
ulvöspinel
Fe_2TiO_4
pseudobrookite
Fe_2TiO_5
titanohematite
titanomagnetite
FeO
wüstite
Fe_3O_4
magnetite
Fe_2O_3
hematite (α)
maghemite (γ)

Fig. 12.2 Ternary compositional diagram of the iron–titanium oxide solid solution magnetic minerals.

Source: M. W. McElhinny, *Paleomagnetism and Plate Tectonics*, Cambridge: Cambridge University Press, 1973.

The most important magnetic minerals belong to the titanomagnetite and titanohematite solid-solution series. The minerals of a third series, pseudobrookite, are paramagnetic at room temperature. They are quite rare and are of negligible importance in rock magnetism.

12.1.3.1 The Titanomagnetite Series

Titanomagnetite is a family of iron oxide minerals described by the general formula $Fe_{3-x}Ti_xO_4$ ($0 \leq x \leq 1$). These minerals have a cubic, inverse spinel structure and exemplify a solid-solution series in which ionic replacement of two Fe^{3+} ions by one Fe^{2+} and one Ti^{4+} ion can take place. The compositional parameter x expresses the relative proportion of titanium in the unit cell. The end members of the solid-solution series are magnetite (Fe_3O_4), which is a typical strongly magnetic ferrite, and ulvöspinel (Fe_2TiO_4), which is antiferromagnetic at very low temperature but is paramagnetic at room temperature. An alternative form of the general formula is $x\,Fe_2TiO_4 \cdot (1-x)\,Fe_3O_4$. Written in this way, it is apparent that the compositional parameter x describes the molecular fraction of ulvöspinel. As the proportion of titanium increases, so does the cell size, which weakens the molecular field. This causes the Curie temperature θ and spontaneous magnetization M_s of the titanomagnetite to decrease.

Magnetite is one of the most important ferrimagnetic minerals. It has a strong spontaneous magnetization ($M_s = 4.8 \times 10^5$ A m^{-1}) and a Curie temperature of 578°C. Because of the high value of M_s, magnetite grains can have a strong shape anisotropy. The magnetic susceptibility is the strongest of any naturally occurring mineral ($k \approx 1$–10 SI). For many sedimentary and igneous rocks the magnetic susceptibility is proportional to the magnetite content.

Maghemite (γ–Fe_2O_3) can be produced by low-temperature oxidation of magnetite and commonly occurs as oxidized rims and fissures in magnetite grains. It is a strongly magnetic mineral ($M_s \approx 3.85 \times 10^5$ A m^{-1}). However, it is metastable and reverts to hematite (α–Fe_2O_3) when heated in air above 300–350°C. The oxidation of titanomagnetite at low temperature results in a "titanomaghemite" solid-solution series.

Titanomagnetite is responsible for the magnetic properties of oceanic basalts. The basaltic layer of the oceanic crust is the main origin of the marine magnetic anomalies that are of vital importance to modern plate tectonic theory. The magnetic properties of the 0.5 km thick basaltic layer are due to the presence of very fine-grained titanomagnetite (or titanomaghemite, depending on the degree of ocean-floor weathering). The molecular fraction (x) of Fe_2TiO_4 in titanomagnetite in oceanic basalts is commonly around 0.6.

12.1.3.2 The Titanohematite Series

The minerals of the titanohematite solid-solution series are also variously referred to as "hemoilmenite," "hematite-ilmenite" or "ilmenohematite." They have the general formula $Fe_{2-x}Ti_xO_3$. The unit cell has rhombohedral symmetry. Ionic substitution is the same as for titanomagnetite, and the compositional parameter x has the same implications for the titanium content of the unit cell. The end members of the solid-solution series are hematite (Fe_2O_3) and ilmenite ($FeTiO_3$). The chemical formula can be written in the alternative form $x\ FeTiO_3 \cdot (1-x)\ Fe_2O_3$, where x represents the molecular fraction of ilmenite. As in the case of titanomagnetite, the cell size increases and the Curie point decreases as the titanium content increases. The Curie point of hematite is 675°C, while ilmenite is antiferromagnetic at low temperature and paramagnetic at room temperature. For titanium contents $0.5 < x < 0.95$, titanohematite is ferrimagnetic and for $x < 0.5$ it exhibits parasitic ferromagnetism.

The end member hematite ($\alpha\text{–}Fe_2O_3$) is an important magnetic mineral in paleomagnetism. Its magnetic properties arise from parasitic ferromagnetism (Appendix C). It has a weak spontaneous magnetization ($M_s \approx 2.2 \times 10^3$ A m^{-1}) and a strong uniaxial magnetocrystalline anisotropy ($K_u \approx 10^3$ J m^{-3}). Hematite is important because of its common occurrence and its high magnetic and chemical stability. It often occurs as a secondary mineral, formed by oxidation of a precursor mineral (e.g., magnetite) or by precipitation from fluids passing through a rock.

12.1.4 Other Ferrimagnetic Minerals

Although the iron–titanium oxides are the dominant magnetic minerals, rocks frequently contain other minerals with ferromagnetic properties. Pyrite (FeS_2) is a very common sulfide mineral, especially in sedimentary rocks, but it is paramagnetic and cannot carry a remanent magnetization. It does not contribute directly to the paleomagnetic properties of rocks, but it may be a source for the formation of goethite or secondary magnetite.

Pyrrhotite is a common sulfide mineral, which can form authigenically or during diagenesis in sediments; it is ferrimagnetic in certain compositional ranges. It is non-stoichiometric (i.e., the numbers of anions and cations in the unit cell are unequal) with the formula $Fe_{1-x}S$. The parameter x refers to the proportion of vacancies among the cation lattice sites and is limited to the range $0 < x < 0.14$. Pyrrhotite has a pseudohexagonal crystal structure and would be antiferromagnetic but for the presence of the cation vacancies. The Néel temperature at which the fundamental antiferromagnetism disappears is around 320°C. Pyrrhotite with the formula Fe_7S_8 is ferrimagnetic, with a Curie temperature close to the Néel temperature and a strong spontaneous magnetization of around 10^5 A m^{-1} at room temperature. The magnetocrystalline anisotropy restricts the easy axis of magnetization to the hexagonal basal plane at room temperature.

Greigite is an iron sulfide with the chemical formula Fe_3S_4. It is a ferrite with an inverse spinel structure like that of magnetite, but with the oxygen atoms replaced by sulfur atoms. It has a comparatively strong spontaneous magnetization ($M_s \approx 1.2 \times 10^5$ A m^{-1}) and exhibits magnetic hysteresis with coercivities similar to magnetite. Like pyrrhotite, it forms authigenically and the time at which it forms is indeterminate. Some studies have associated it with early diagenesis, others have found that both greigite and pyrrhotite carry late diagenetic magnetizations. Greigite is unstable thermally, decomposing at temperatures above 200°C to form daughter products that, on further heating, facilitate the growth of other magnetic minerals. Consequently, the Curie temperature of greigite cannot be measured directly. Despite its instability, greigite can be a carrier of remanent magnetization in lacustrine and marine sediments, and thus can have a significant role in environmental magnetism.

Goethite, an iron oxyhydroxide (FeOOH), is another common authigenic mineral in sediments. Like hematite, goethite is antiferromagnetic, but has a weak parasitic ferromagnetism. It has a weak spontaneous magnetization ($M_s \approx 2 \times 10^3$ A m^{-1}) and a very high coercivity (with maximum values in excess of 5 T) and a low Curie point around 100°C or less. It is thermally unstable relative to hematite under most natural conditions, and decomposes on heating above about 350°C. It is a common (and paleomagnetically undesirable) secondary mineral in limestones and other sedimentary rocks.

12.1.5 Identification of Ferrimagnetic Minerals

It is often difficult to identify the ferrimagnetic minerals in a rock, because their concentration is very low, especially in sedimentary rocks. If the rock is coarse-grained, ferrimagnetic minerals may be identified optically among the opaque grains by studying polished sections in reflected light. However, many rocks that are paleomagnetically important (e.g., basaltic lava, pelagic limestone) have a very fine grain size. Optical examination, even with a microscope, is usually unable to identify the ferrimagnetic mineral. Examination with a transmission electron microscope (TEM) using energy dispersive X-ray spectroscopy provides a direct method of identifying mineral grains in powders that have been separated from a rock or sediment. The mineral responsible for the remanent magnetization in a rock may also be identified by its Curie temperature and hysteresis properties.

The Curie temperature is measured using a form of balance in which a strong magnetic field gradient exerts a force on the sample that is proportional to its magnetization. The sample is heated and the change of magnetic force (i.e., sample magnetization) is observed with increasing temperature. When the Curie point is reached, the ferromagnetic behavior disappears; at higher temperatures the sample is paramagnetic. The Curie point is diagnostic of many minerals. However, heating can cause unstable or metastable minerals to break down (e.g., due to oxidation) before the Curie temperature is reached. An example of Curie temperature analysis on an extract of magnetic minerals from a pelagic limestone (Fig. 12.3) shows the presence of goethite ($\theta_g \sim 100^\circ$C, sample SR 3A), magnetite ($\theta_m \sim 570^\circ$C, all samples) and hematite ($\theta_h \sim 650^\circ$C, sample SR 11). As in this case, Curie point analysis of most rocks requires that the ferrimagnetic minerals be extracted or concentrated. Extraction is sometimes difficult, and often it is not certain that the extract is representative of the rock as a whole. This is also a drawback of optical methods;

microscopic examination only allows description of grains that are large enough to be viewed, while transmission electron microscopy is also based on a magnetic separate.

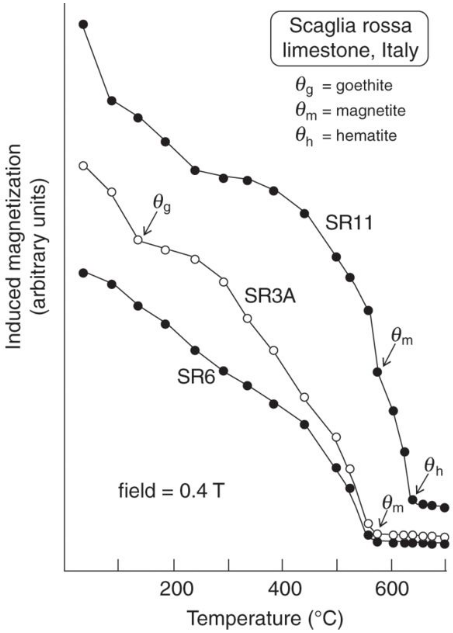


Fig. 12.3 Identification of the ferromagnetic minerals in a pelagic limestone by determination of their Curie temperatures in concentrated extracts (after Lowrie and Alvarez, 1975).

12.1.6 Grain Size Dependence of Ferrimagnetic Properties

The ferromagnetic properties of metals and ferrites vary sensitively with grain size. Consider an assemblage of uniformly magnetized grains of a ferrimagnetic mineral characterized by a spontaneous magnetization M_s and uniform grain volume V. Let the spontaneous magnetization of any grain be oriented parallel to an easy direction of magnetization (crystalline or shape-determined), defined by the anisotropy energy K_u per unit volume. The energy that keeps the magnetization parallel to the easy direction in a grain is equal to VK_u. Thermal energy, proportional to the temperature, has the effect of disturbing this alignment. At temperature T the thermal energy is equal to kT, where k is Boltzmann's constant ($k = 1.381 \times 10^{-23}$ J K^{-1}). At any instant in time there is a chance that thermal energy will deflect the magnetic moment of any grain in the assemblage away from its easy direction. Progressively, the net magnetization of the material (the sum of all the magnetic moments of the numerous magnetic grains) will be randomized by the thermal energy, and the magnetization will be observed to decay. If the initial magnetization of the assemblage is M_{ro}, after time t it will decrease exponentially to $M_r(t)$, according to

$$M_r(t) = M_{r0} \exp\left(\frac{-t}{\tau}\right). \tag{12.6}$$

In this equation τ is known as the *relaxation time* (Box 12.1). If the relaxation time is long, the exponential decrease in Eq. (12.6) is slow and the magnetization is stable. The time τ depends on properties of the grain and is given by the equation

$$\tau = \frac{1}{\nu_0} \exp\left(\frac{VK_u}{kT}\right). \tag{12.7}$$

The constant ν_0 is related to the lattice vibrational frequency and has a very large value ($\approx 10^8$–10^{10} s^{-1}). The value of K_u depends on whether the easy direction of the magnetic mineral is determined by the magnetocrystalline anisotropy or the magnetostatic (shape) anisotropy.

Box 12.1 Magnetic Relaxation

Relaxation behavior is characterized by an exponential change with time from a state of elevated energy to a state of lower energy. The most familiar example is radioactive decay (Section 8.3.1), but the magnetizations of natural materials also exhibit relaxation behavior, as explained in Section 12.1.5.

In the absence of an external field, the easy direction of magnetization of a single-domain magnetic grain with volume V and anisotropy energy K_u per unit volume is determined by the anisotropy energy VK_u. The probability that the grain's thermal energy kT can overcome this energy barrier and allow the grain magnetization to change direction is determined by the Maxwell–Boltzmann distribution of energies, and is proportional to $\exp(-VK_u/kT)$. The probability per unit time, λ, of a magnetic moment, μ, changing to a different easy axis is given by the *Arrhenius equation* for a thermally activated process

$$\lambda = C\exp\left(-\frac{VK_u}{kT}\right). \tag{1}$$

The parameter C is called the frequency factor; here it is the lattice vibration frequency, ν_0. At a particular temperature, all parameters on the right side of Eq. (1) are constant and so the probability λ per unit time of a magnetic moment changing to a different easy axis is constant.

Suppose an assemblage of identical non-interacting single domain particles, of which N_1 are magnetized initially in one direction (state 1) and N_2 in the opposite direction (state 2). The net magnetization is proportional to $(N_1 - N_2)$. Assuming a constant probability λ per unit time, the number of particles $\mathrm{d}N_1$ that change from state 1 to state 2 is proportional to the time interval $\mathrm{d}t$ and to the number of grains N_1 in state 1. It is given by

$$\mathrm{d}N_1 = -\lambda N_1 \mathrm{d}t, \tag{2}$$

where the negative sign indicates a reduction in N_1. Similarly, the number changing in the opposite sense from state 2 to state 1 is

$$\mathrm{d}N_2 = -\lambda N_2 \mathrm{d}t. \tag{3}$$

The net change in magnetization is therefore

$$\mathrm{d}M = \mu \mathrm{d}N_1 - \mu \mathrm{d}N_2 = \mu \mathrm{d}(N_1 - N_2), \tag{4}$$

$$\mathrm{d}M = -\lambda\mu(N_1 - N_2)\mathrm{d}t = -\lambda M \mathrm{d}t. \tag{5}$$

The solution of this differential equation is

$$M = M_0 \exp(-\lambda t) = M_0 \exp\left(-\frac{t}{\tau}\right), \tag{6}$$

where τ, the relaxation time, is the inverse of λ in Eq. (1):

$$\tau = \frac{1}{v_0} \exp\left(\frac{VK_u}{kT}\right). \tag{7}$$

Equations (6) and (7) are important in paleomagnetism. The strong anisotropy of fine-grained ferromagnetic minerals can result in very long relaxation times, and consequently the magnetizations of fine-grained rocks can be extremely stable over geological lengths of time.

This theory applies only to very small grains that are uniformly magnetized. Fine-grained ferrimagnetic minerals are, however, very important in paleomagnetism and rock magnetism. The very finest grains, smaller than a critical size, exhibit an unstable type of magnetic behavior called *superparamagnetism*, with relaxation times typically less than 100 s. Above the critical size the uniformly magnetized grain is very stable and is called a single-domain grain.

12.1.6.1 Superparamagnetism

In a ferromagnetic material the strong molecular fields keep the atomic spin magnetic moments uniformly aligned with each other; the grain anisotropy requires this spontaneous magnetization to lie parallel to an “easy” direction. If the temperature is high enough, thermal energy (kT) may exceed the anisotropy energy (VK_u) but still be too small to overpower the spontaneous magnetization. The thermal energy then causes the entire magnetic moment of the grain to fluctuate coherently in a manner similar to paramagnetism (the theory of which applies to individual atomic magnetic moments). The grain magnetization has no stable direction, and the behavior is said to be *superparamagnetic*. It is important to note that superparamagnetic grains themselves are immobile; only their uniform magnetization fluctuates relative to the grain.

Whether the state of a ferrimagnetic grain is unstable and superparamagnetic or stable and single-domain depends on the grain size, the magnetic anisotropy energy, and the temperature. If the grain volume V is very small, unstable magnetic behavior due to superparamagnetism becomes likely. Magnetite and hematite grains finer than about 0.03 μm in diameter are superparamagnetic at room temperature.

12.1.6.2 Single-Domain Particles

When the anisotropic magnetic energy (VK_u) of a grain is greater than the thermal energy (kT), the spontaneous magnetization direction favors one of the easy directions. The entire grain is uniformly magnetized as a *single domain*. This situation occurs in very fine grains of ferrimagnetic minerals.

In magnetite K_u is the magnetostatic energy related to the particle shape. The theoretical range of single-domain sizes in magnetite is narrow, from about 0.03 to 0.1 μm in equant grains and up to about 1 μm in elongate grains (Fig. 12.4). In hematite K_u is the large magnetocrystalline anisotropy energy, and the range of single-domain sizes is larger, from about 0.03 to 15 μm.

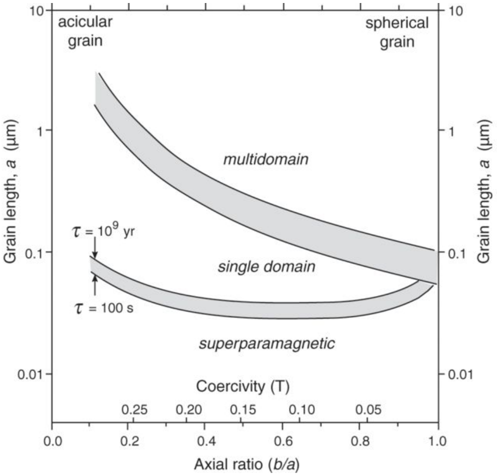


Fig. 12.4 Ranges of grain sizes and shapes for superparamagnetic, single domain, and multidomain magnetic behavior in ellipsoidal magnetite grains. Source: M. E. Evans and M. W. McElhinny, An investigation of the origin of stable remanence in magnetite-bearing igneous rocks, *J. Geomag. Geoelect.* 21, 757–773, 1969.

The magnetization of a single-domain particle is very stable, because to change it requires rotating the entire uniform spontaneous magnetization of the grain against the grain anisotropy, which requires a very strong magnetic field. The magnetic field required to reverse the direction of magnetization of a single-domain grain is called its *coercivity* B_c and is given by:

$$B_c = \frac{2K_u}{M_s}. \tag{12.8}$$

The maximum coercivity of single-domain magnetite is around 0.3 T for needle-shaped elongate grains. The strong magnetocrystalline anisotropy of hematite gives it a higher maximum coercivity, which commonly may exceed 2 T. Because of their stable remanent magnetizations, single-domain particles play an important role in paleomagnetism.

12.1.6.3 Multidomain Particles

Single-domain behavior is restricted to a limited range of grain sizes. When a grain is large enough, the magnetic energy associated with its magnetization becomes too large for the magnetization to remain uniform. This is because the demagnetizing field of a uniformly magnetized grain (Fig. 12.5a) interacts with the spontaneous magnetization and generates a magnetostatic (or self-demagnetizing) energy. To reduce this energy, the magnetization subdivides into smaller, uniformly magnetized units, called *Weiss domains* after P. Weiss, who theoretically predicted domain structure in 1907. In the simplest case the magnetization divides into two, oppositely magnetized domains (Fig. 12.5b). The net magnetization is reduced to zero, and the magnetostatic energy is reduced by about half. Further subdivision (Fig. 12.5c) reduces the magnetostatic energy correspondingly. In a grain with n domains of alternately opposed spontaneous magnetizations the magnetostatic energy is reduced by a factor of $1/n$. Depending on the crystal symmetry of the magnetic mineral, so-called closure domains may develop at the ends of the domains, reducing the magnetostatic energy further.

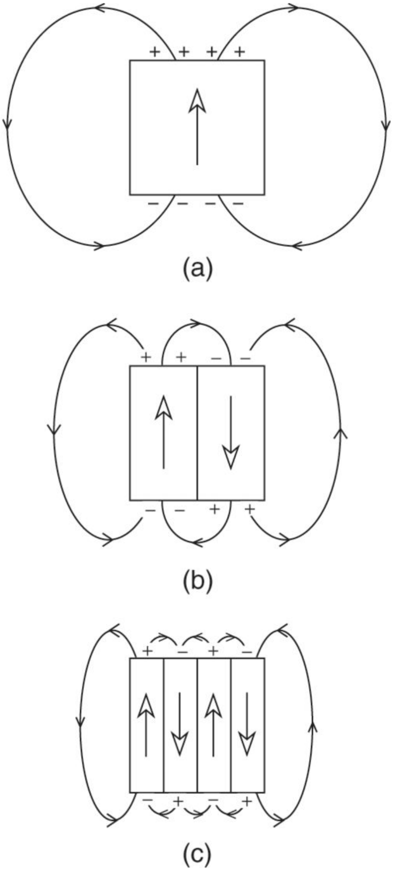


Fig. 12.5 Subdivision of (a) the uniform magnetization of a large grain into (b) two oppositely magnetized magnetic domains and (c) four alternately magnetized domains.

The domains are separated from one another by thin regions, about 0.1 μm thick, that are usually much thinner than the domains and are called *Bloch domain walls*. Within the domain wall the magnetization undergoes small progressive changes in direction from each atom to its neighbor. The crystalline magnetic anisotropy of the material attempts to keep the atomic magnetic spins parallel to favored crystalline directions, while the exchange energy resists any change of direction from parallel alignment by the molecular field. The energy of these competing effects is expressed as a *domain wall energy* associated with each unit area of the wall.

The magnetization of a *multidomain* grain can be changed by moving the position of a domain wall, which causes some domains to increase in size and others to decrease. A large multidomain grain may contain many easily movable domain walls. Consequently, it is much easier to change the magnetization of a multidomain grain than of a single-domain grain. As a result multidomain grains are less stable carriers of remanent magnetization than single-domain grains and coarse-grained rocks are often unsuitable for paleomagnetism.

12.1.6.4 Pseudo-Single-Domain Particles

The transition between single-domain and multidomain behavior depends on grain shape as well as size (Fig. 12.4). Magnetite grains finer than 0.1 μm in size are in a single-domain state, which in needle-shaped grains may persist up to about 1 μm in length. Magnetite grains larger than a few μm in diameter are multidomain. Between these extremes a grain is physically too small to accommodate a domain wall (thickness ~0.1 μm) but too large to have a uniform magnetization. The intermediate range of grain sizes is referred to as pseudo-single domain (PSD), because the intensity, coercivity, and stability of the magnetization are comparable to the properties of single-domain grains.

The grain-size range for stable single-domain behavior is so narrow that a large proportion of the magnetizations of rocks must be carried by PSD grains. The nature of the magnetization in a PSD grain is a matter of conjecture due to limited ability to observe them. However, computer models of the configurations of the magnetic spins in PSD grains have brought a fresh interpretation of their internal state.

Physical studies of the orientations of magnetic spins in media with restricted dimensions (for example, within ultrathin magnetic films), have led to the development of *micromagnetic* models of spin orientations. According to these models, PSD grains do not have uniform magnetizations; instead, the magnetic spins orient so as to minimize the directional change between neighboring spins, and thereby the exchange energy between atoms. At grain surfaces the spins bend away from the surface so as to intersect it at the shallowest angle possible, in this way reducing the magnetostatic energy. Geometrically this results in a vortex-like configuration of magnetic spins in tiny magnetic particles. Micromagnetic models have shown that PSD particles with vortex-like spin orientations can have high coercivity and directional stability. PSD grains are therefore able to act as reliable recorders of paleomagnetic directions.

12.1.6.5 First-Order Reversal Curves

The domain state of the magnetic grains in a rock or sediment can be determined by *first order reversal curve* (FORC) analysis. This technique is based on the magnetic hysteresis displayed by ferromagnetic particles. A sample is first magnetized to saturation as in Appendix C, Fig. C3. The magnetizing field is then reversed to a value B_r; from this point a new magnetizing curve is traced out as shown by the dashed curve in Fig. C3. In a FORC analysis about 100 of these magnetizing curves are measured (Fig. 12.6a), with different starting fields B_r, so that the full area of the hysteresis diagram is mapped. FORC analysis reduces the observed curves to a distribution of two fields: the coercivities B_c of the grains in the rock, and the interaction fields B_i between the grains. These are plotted against each other in a FORC diagram to reveal the size and character of the magnetic grains.

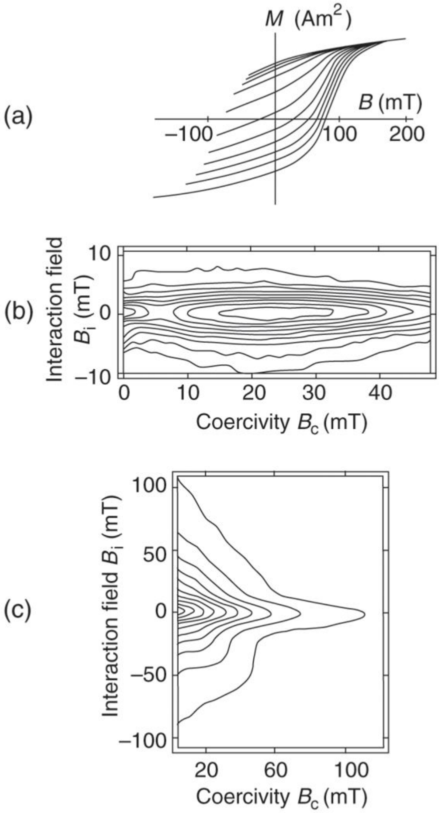


Fig. 12.6 First-order reversal (FORC) analysis of magnetic grain size. (a) Family of magnetization curves with different starting field, B_r. Typical FORC diagrams for (b) a sample with non-interacting single-domain particles, and (c) multidomain particles. Source: A. P. Roberts, C. R. Pike and K. L. Verosub, First-order reversal curve diagrams: A new tool for characterizing the magnetic properties of natural samples, *J. Geophys. Res.*, 105, 28461–28475, 2000.

This is illustrated by two extreme situations. An assemblage of non-interacting single-domain grains has negligible interaction fields but possesses a spectrum of coercivities. On a FORC diagram the coercivity spectrum maps as a narrow ridge close to the horizontal axis where B_i is zero (Fig. 12.6b). On the other hand, multidomain particles have low coercivities and, because each domain is dependent on its neighbors, a broad distribution of interaction fields. Thus, the FORC diagram of a multidomain sample shows contours that cluster near the origin of the plot and extend along the vertical B_i axis (Fig. 12.6c). FORC analysis is a powerful way of resolving the magnetic character of rocks with combinations of differently sized grains, including the superparamagnetic and PSD vortex states.

12.1.7 Remanent Magnetizations in Rocks

Although a rock usually contains only a small concentration of ferrimagnetic minerals, they enable it to acquire a remanent magnetization (or *remanence*). The remanence of a rock before laboratory analysis is called its *natural remanent magnetization* (NRM). It may consist of several components acquired in different ways and at different times. The geologically important types of remanence are acquired at known times in the rock's history, such as at the time of its formation or subsequent alteration. The remanence of a rock can be very stable against change; the high coercivity of the ferrimagnetic mineral, especially in fine grains, assures preservation of the magnetic signal during long geological epochs.

A remanence acquired at or close to the time of formation of the rock is called a *primary* magnetization; a remanence acquired at a later time is called a *secondary* magnetization. Examples of primary remanence are thermoremanent magnetization, which an igneous rock acquires during cooling, and the remanent magnetizations acquired by a sediment during or closely after deposition. Secondary remanences may be caused by chemical change of the rock during diagenesis or weathering, or by the sampling and laboratory procedures.

12.1.7.1 Thermoremanent Magnetization

The most important type of remanent magnetization in igneous (and high-grade metamorphic) rocks is *thermoremanent magnetization* (TRM). Igneous rocks solidify at temperatures above 1000°C. Below this temperature the mineral grains are fixed in a rigid matrix. The grains of a ferrimagnetic mineral are well above their Curie temperature (Fig. 12.7a), which in magnetite is 578°C and in hematite is 675°C. The molecular field is not active and the individual atomic magnetic moments are free to fluctuate chaotically; the magnetization is therefore paramagnetic at high temperatures.

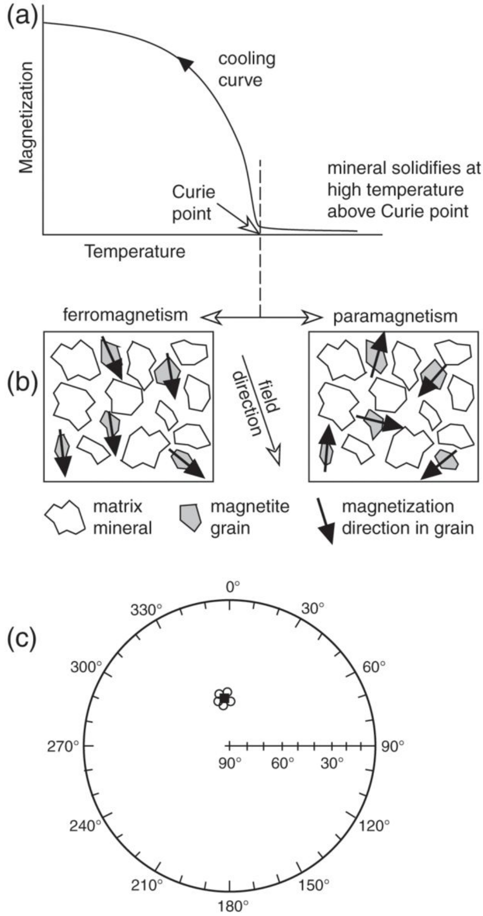

(a)
Magnetization
cooling curve
Curie point
mineral solidifies at high temperature above Curie point
Temperature
ferromagnetism
paramagnetism
(b)
field direction
matrix mineral
magnetite grain
magnetization direction in grain
(c)
0°
30°
60°
90°
120°
150°
180°
210°
240°
270°
300°
330°
90°
60°
30°

Fig. 12.7 (a) On cooling through the Curie temperature the magnetic state of magnetite grains changes from paramagnetism to ferromagnetism. (b) Below the Curie point the magnetizations inside the magnetite grains become blocked along easy directions of magnetization close to the field direction. (c) Stereogram showing directions of TRM (open circles) in a basaltic lava flow on Mt. Etna (Sicily) and the direction of the geomagnetic field (solid square) at the time of the eruption (data from Chevallier, 1925).

As the rock cools, the temperature eventually passes below the Curie temperature of the ferrimagnetic grains and a spontaneous magnetization appears. In single-domain grains the relaxation time of the grain magnetization is governed by Eq. (12.7), which can be modified by writing the anisotropy energy density K_u in terms of the spontaneous magnetization M_s and coercivity B_c of the grain. This gives the relaxation time as follows:

$$\tau = \frac{1}{\nu_0}\exp\left(\frac{VM_s\,B_c}{2kT}\right). \tag{12.9}$$

At high temperature the thermal energy (kT) is larger than the magnetic energy ($VM_sB_c/2$) and the magnetization is unstable. Although the individual atomic magnetic moments are forced by the molecular field to act as coherent units, the grain magnetizations are superparamagnetic. As the rock cools further, the spontaneous magnetization and the magnetic anisotropy energy K_u of the grain increase. Eventually the temperature passes below a value at which the randomizing thermal energy is no longer greater than the magnetic anisotropy energy. The spontaneous magnetization then becomes “blocked” along an easy direction of magnetization of the grain. In the absence of an external magnetic field the grain magnetic moments are randomly oriented. Once the grain cools below its *blocking temperature* in a magnetic field, the grain magnetic moment is “blocked” along the easy axis that is closest to the direction of the field at that time (Fig. 12.7b). The alignment of grain magnetic moments with the field is neither perfect nor complete; it is a statistical preference. This means that in an assemblage of grains, more of the grains have their magnetic moments aligned close to the field direction than any other direction.

It is important to note that the mineral grains themselves are immobile throughout the process of acquisition of TRM. Only the internal magnetizations of the grains can change direction and eventually become blocked. The blocking temperatures of TRM are dependent on the grain size, grain shape, spontaneous magnetization, and magnetic anisotropy of the ferrimagnetic mineral. If the rock contains a wide range of grain sizes and perhaps more than a single magnetic mineral, there may be a broad spectrum of blocking temperatures. Maximum blocking temperatures may range as high as the Curie point, and the spectrum can extend to below ambient temperature. If the magnetic field is applied only while the rock is cooling through a limited temperature range, only grains with blocking temperatures in this range are activated, and a partial TRM (or pTRM) results.

TRM is a very stable magnetization that can exist unchanged for long intervals of geological time. The ability of TRM to record accurately the field direction is demonstrated by the results from a lava that erupted on Mt. Etna at a time for which a record of the magnetic field direction was available from observatory data. The directions of the TRM in the lava samples are the same as the direction of the ambient field (Fig. 12.7c).

12.1.7.2 Sedimentary Remanent Magnetizations

The acquisition of *depositional remanent magnetization* (DRM) during deposition of a sediment takes place at constant temperature. Magnetic and mechanical forces compete to produce a physical alignment of detrital ferrimagnetic particles. During settling through still water these particles are oriented by the ambient magnetic field in the same way that it orients a compass needle. The particles become aligned statistically with the Earth's magnetic field (Fig. 12.8). The action of mechanical forces may at times spoil this alignment. Water currents cause hydromechanical forces that disturb the alignment during settling, giving rise to a *declination error*. On contact with the bottom of the sedimentary basin, the mechanical force of gravity rolls the particle into a stable attitude, causing an *inclination error*. The pressure of overlying sediment during deep burial results in compaction, which can produce further inclination shallowing. The DRM is finally fixed in sedimentary rocks during diagenesis.

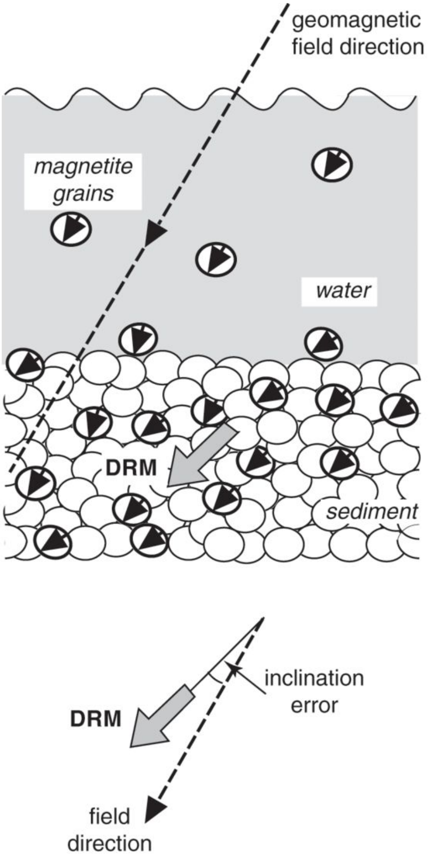

geomagnetic
field direction
magnetite
grains
water
DRM
sediment
DRM
inclination
error
field
direction

Fig. 12.8 Acquisition of depositional remanent magnetization (DRM) in a sediment; gravity causes an inclination error between the magnetization and field directions.

Source: E. Irving and A. Major, Post-depositional detrital remanent magnetization in a synthetic sediment, *Sedimentology*, 3, 135–143, 1964.

A modified form of *post-depositional remanence* (pDRM) is important in fine-grained sediments. A waterlogged slurry forms at the sediment–water interface. Fine-grained magnetic minerals in the water-filled pore spaces in the sediment are partially in suspension and may be reoriented by the magnetic field, if they are free enough to move (Fig. 12.9a). The energy for this motion of the particles is obtained from the Brownian motion of the water molecules, which continuously and randomly collide with the particles in the pore spaces. Large particles are probably in contact with the surrounding grains and are unaffected by collisions with the water molecules. Very fine particles that are virtually floating in the pore spaces may acquire enough freedom from the Brownian agitation to align statistically with the Earth's magnetic field. Laboratory experiments have established that the direction of pDRM is an accurate record of the depositional field, without inclination error (Fig. 12.9b). The pDRM is acquired later than the actual time of sedimentation, and is fixed in the sediment during compaction and de-watering at a depth of ~10 cm. This may represent a lock-in time delay of 100 years in lacustrine sediments or as much as 10,000 years in pelagic marine sediments. In the latter case the lock-in delay is usually less than the errors involved in locating paleontological stage boundaries. The pDRM process is particularly effective in fine-grained sediments containing strongly magnetic magnetite grains. For example, pDRM is the most important mechanism of primary magnetization in pelagic limestones. Compaction may cause a flattening of the inclination under some conditions.

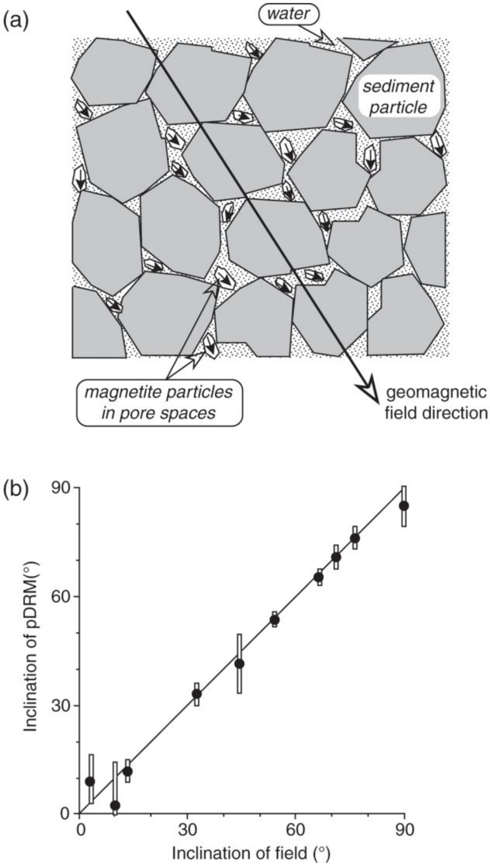


Fig. 12.9 (a) Post-depositional remanent magnetization (pDRM) is acquired by reorientation of ferromagnetic grains in the pore spaces of a deposited sediment. (b) Comparison of the pDRM inclination with the field inclination in a redeposited deep-sea sediment (after Irving and Major, 1964).

Bioturbation may mix the sediment, typically to a depth of about 10 cm in pelagic sediments. This affects the positions of stratigraphic marker levels. First-occurrence datum levels of fossils are carried deeper, and last occurrences are carried higher than the true stratigraphic levels. Agitation of bioturbated sediment by the burrowing organisms assists the Brownian motion of the magnetic particles. Under these conditions the pDRM is acquired at the base of the bioturbated zone.

12.1.7.3 Chemical Remanent Magnetization

Chemical remanent magnetization (CRM) is usually a secondary form of remanence in a rock. It occurs when the magnetic minerals in a rock suffer chemical alteration or when new minerals form authigenically. An example is the precipitation of hematite from a goethite precursor or from iron-saturated fluids that pass through the rock. The magnetic minerals may also experience diagenetic modification or oxidation by weathering, which usually happens on the grain surface and along cracks (Fig. 12.10). The growth of a new mineral (or the alteration of an existing one) involves changes in grain volume V, spontaneous magnetization M_s, and coercivity B_c. The chemical change affects the relaxation time of the magnetization, according to Eq. (12.9). The grains eventually grow through a critical volume, at which the magnetization becomes blocked. The new CRM is acquired in the direction of the ambient field during the chemical change, and so it is younger than the host rock. It has stable magnetic properties similar to TRM. A common example is the formation of pigmentary hematite during diagenesis or weathering. Hematite that originates in this way typically carries a secondary remanent magnetization.

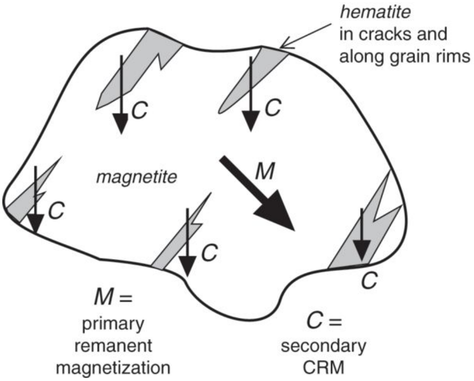


Fig. 12.10 Acquisition of chemical remanent magnetization (CRM) accompanies the diagenetic modification or oxidation by weathering of magnetic minerals; this often happens on the grain surface and along cracks.

12.1.7.4 Isothermal Remanent Magnetization

Isothermal remanent magnetization (IRM) is induced in a rock sample by placing it in a magnetic field at constant temperature. For example, rock samples are exposed to the magnetic fields of the sampling equipment and to other magnetic fields during transport to the laboratory. A common technique of rock magnetic analysis consists of deliberately inducing IRM in a known magnetic field produced in a large coil or between the poles of an electromagnet. The magnetic moments within each grain are partially aligned by the applied field. The degree of alignment depends on the field strength and on the resistance of the magnetic mineral to being magnetized, i.e., its coercivity. After removing the sample from the applied field an IRM remains in the rock (Appendix C, Fig. C3). This artificially induced remanent magnetization is useful in various rock magnetic applications.

If the rock sample is placed in progressively stronger fields, the IRM increases to a maximum value called the saturation IRM, which is determined by the type and concentration of the magnetic mineral. The shape of the progressive acquisition curve and the field needed to reach saturation IRM depend on the coercivities of the magnetic minerals in the rock, which are well known (Table 12.1).

Table 12.1 Maximum coercivities and blocking temperatures for some common ferromagnetic minerals

Ferromagnetic mineral	Maximum coercivity (T)	Maximum blocking temperature (°C)
Magnetite	0.3	575
Maghemite	0.3	≈ 350
Titanomagnetite ($Fe_{3-x}Ti_xO_4$):		
$x = 0.3$	0.2	350
$x = 0.6$	0.1	150
Pyrrhotite	0.5–1	325
Hematite	1.5–5	675
Goethite	> 5	80–120

12.1.7.5 Other Remanent Magnetizations

The magnetization of a rock that contains magnetic minerals with unstable magnetic moments may change in direction and intensity as time goes on. The time-dependent growth or decay of the magnetization changes logarithmically with time and is called a *viscous remanent magnetization* (VRM). Its direction in rocks is often parallel to the present-day field direction, which can be useful in its identification. It is a secondary magnetization and it can mask the presence of geologically interesting stable components. Techniques of progressive demagnetization (Section 12.2.3.2) are used to "clean" the magnetization of a rock of undesirable components such as VRM. The "magnetic cleaning" randomizes the orientations of unstable magnetizations in a rock, and is carried out in a field-free space.

If the randomization is done by means of an alternating magnetic field and a residual magnetic field is present in the equipment, the magnetic moments are not randomized but have a statistical preference for the direction of the residual field. The sample acquires an *anhysteretic remanent magnetization* (ARM), which can prevent isolation of the primary magnetization component in a paleomagnetic investigation. An ARM may also be induced deliberately, to compare its characteristics with the natural remanent magnetization.

The history of a rock sample is reflected in its magnetization. Stress produces strains in a mineral, changing the symmetry of its crystal lattice. This results in magnetostrictive energy (Appendix D), which is a source of magnetic anisotropy that can modify the easy directions of magnetization in some magnetic minerals. Nonhydrostatic stress in the presence of a magnetic field may produce a *piezoremanent magnetization* (PRM) that is added to, or may replace, any preexisting remanent magnetization. Consequently, elastic or plastic stress can overprint the original magnetization in rocks that have been deeply buried, or which are tectonically deformed.

A related type of magnetization has been observed in terrestrial and lunar rocks that have been shocked by meteoritic impact. The event causes very high local stress of short duration, and gives rise to a *shock remanent magnetization* (SRM) by a similar mechanism to PRM. The collision of a meteorite with another planetary body imposes a characteristic shock texture on the impacted rocks, so that SRM and PRM are not exactly equivalent. It is possible that SRM has played a role in the origin of anomalous crustal magnetizations on the Moon and Mars, which have suffered extensive meteoritic bombardment.

12.1.8 Environmental Magnetism

Since the mid-1980s, rock magnetic properties have found important new applications in environmental research, where they can serve as tracers for pollution or as indicators of past climates. Heavy metals from industrial processes enter the environment, contaminating both the ground surface and the water in lakes, rivers, and the groundwater. Magnetic minerals such as magnetite and hematite accompany more toxic elements in the emissions. Motor-vehicle exhaust emissions also pollute the environment by loading it with heavy metals as well as sub-micron-sized particles of soot, a fraction of which consists of nanoparticle-sized magnetic minerals. Rock magnetic parameters and investigative techniques provide ways of characterizing these particles and of describing their regional distribution and concentration.

The magnetic mineral in a sediment or soil usually consists of hematite, magnetite, maghemite, or an iron sulfide. It may be modified by the climatic conditions at or after deposition. This affects magnetic properties of the sediment that depend on grain size and mineral composition, such as magnetic susceptibility, coercivity, and various remanent magnetizations. These magnetic parameters can act as proxies for paleoclimatic change. Numerous investigations have been devoted to developing the potential of magnetic methods for analyzing a paleoclimatic problem, as illustrated by the following example.

Great thicknesses of loess sediments up to 100 m in thickness occur in central China. The loess are very fine-grained (10–50 μm) wind-blown sediments that are deposited in cold, dry conditions. Alternating with the beds of loess are layers of *paleosol*, the name given to fossil soils. These form by conversion of the loess to soils (a process called *pedogenesis*), which takes place during interglacial periods of warmer, humid conditions. In turn, later loess deposits bury the soils. The alternation of loess and paleosol is interpreted as a record of past climatic variations. The magnetic susceptibility correlates with the lithology of the loess–paleosol sequences (Fig. 12.11). In the loess beds the susceptibility measures around 25 SI units, but in the paleosol layers the values exceed 200 SI units. Rock magnetic analysis identified the dominant magnetic mineral in both lithologies as magnetite. The higher susceptibility of the paleosols is thus due to higher concentrations of magnetite. The susceptibility variation correlates very well with the oxygen isotope stratigraphy measured in deep-sea sediments at site 677 of the Ocean Drilling Program (ODP). The oxygen isotope ratio (Box 8.1) is a record of climatic variation, with the most negative values corresponding to warmer temperature. It has been dated by correlation with the magnetic reversal record in marine sediments. The correlation in Fig. 12.11 provides a way of dating the loess–paleosol sequence and shows that the paleosols were formed under warmer conditions than the loess. Possibly the change to a warmer, humid climate encouraged the *in-situ* formation of a new phase of magnetite in the paleosols, with a consequent increase in susceptibility.

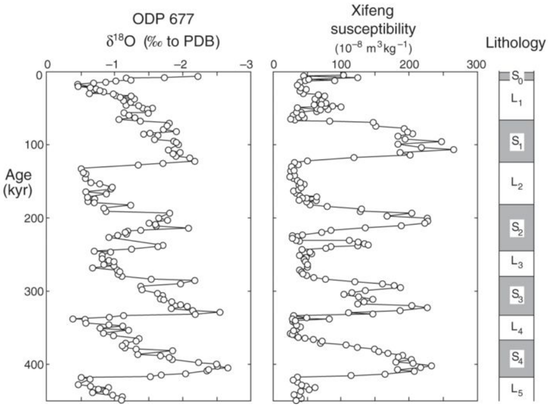


Fig. 12.11 Comparison of the magnetic susceptibility variation in a section of loess and paleosols at Xifeng, China, with the oxygen isotope record in marine sediments cored at ODP site 677.

Source: M. E. Evans and F. Heller, *Environmental Magnetism: Principles and Applications of Enviromagnetics*, New York: Academic Press, 2003.

The new phase of magnetite in the paleosols originates by a chemical process. However, magnetite can be produced by the action of bacteria in sediments deposited in lakes and in the sea. This is called a *biogenic* process and the bacteria that can produce magnetite are called *magnetotactic bacteria*. The magnetite forms as tiny crystallites called *magnetosomes,* less than 0.1 μm in size and thus of single-domain type, which occur as chains of particles enclosed in a membrane. The most common magnetosome mineral is magnetite, but greigite, an iron sulfide with structure similar to magnetite, has also been found. The chain-like assembly imparts a dipole magnetic moment to the bacteria. This is evidently an evolutionary feature, which helps the bacteria to survive. If the sediment is disturbed, the magnetic moment of the bacteria enables them to move along the Earth’s magnetic field lines. The dipping field lines guide them back down into the sediment, where they find nutrition needed for their survival. It was long believed that magnetite in deep-sea sediments was mainly of detrital origin, washed in from the continents and ocean ridges, but magnetite of biogenic origin may be an important component.

Biogenic magnetite has been identified in many unusual settings. For example, submicroscopic magnetite occurs in the brains of dolphins and birds; whether it plays a role in their guidance during migration is unclear. Magnetite of nanometer grain size has been located in the human brain, where it may be related to neurological disorders such as epilepsy, Alzheimer’s disease, and Parkinson’s disease. It also occurs in other human organs. The magnetic properties of the magnetite, its source, and the mechanism of its formation in human tissue are current fields of research.

12.2 Apparent Polar Wander and Tectonics

12.2.1 Rock Magnetization and Continental Drift

A mountain walker using a compass to find his way in the Swiss Alps above the high mountain valley of the Engadine would notice that in certain regions (for example, south of the Septimer Pass) the compass needle shows very large deviations from the north direction. The deflection is due to the local presence of strongly magnetized serpentinites and ultramafic rocks. Early compasses were more primitive than modern versions, but the falsification of a compass direction near strongly magnetic outcrops was known by at least the early nineteenth century. In 1797 Alexander von Humboldt proposed that the rocks in these unusual outcrops had been magnetized by lightning strikes. The first systematic observations of rock magnetic properties are attributed to A. Delesse (1849) and M. Melloni (1853), who concluded that volcanic rocks acquired a remanent magnetization during cooling. After a more extensive series of studies in 1894 and 1895 of the origin of magnetism in lavas, G. Folgerhaiter reached the same conclusion and suggested that the direction of remanent magnetization was that of the geomagnetic field during cooling. By 1899 he had extended his work to the record of the secular variation of inclination in ancient potteries. Folgerhaiter noted that some rocks have a remanent magnetization opposite to the direction of the present-day field. Reversals of polarity of the geomagnetic field were established decisively early in the twentieth century.

In 1922 Alfred Wegener proposed the concept of continental drift, based on years of study of paleoclimatic indicators such as the geographic distribution of coal deposits. At the time, there was no way of explaining the mechanism by which the continents drifted. Only motions of the crust were considered, and the idea of rigid continents plowing through rigid oceanic crust was unacceptable to geophysicists. There was as yet no way to reconstruct the positions of the continents in earlier eras or to trace their relative motions. Subsequently, paleomagnetism was to make important contributions to understanding continental drift by providing the means to trace past continental motions quantitatively.

In the 1950s, several small research groups were engaged in determining and interpreting the directions of magnetization of rocks of different ages in Europe, Africa, North and South America, and Australia. In 1956 S. K. Runcorn put forward the first clear geophysical evidence in support of continental drift. Runcorn compared the directions of magnetization of Permian and Triassic rocks from Great Britain and North America. He found that the paleomagnetic results from the different continents could be brought into harmony for the time before 200 Myr ago by closing the Atlantic Ocean. The evaluation of the scientific data was statistical and at first was regarded as controversial. However, Mesozoic paleomagnetic data were soon obtained from the southern hemisphere that also argued strongly in favor of the continental drift hypothesis. In 1957 E. Irving showed that paleomagnetic data conformed better with geological reconstructions of earlier positions of the continents than with their present-day distribution. Subsequently, numerous studies have documented the importance of paleomagnetism as a chronicle of past motions of global plates and as a record of the polarity history of the Earth’s magnetic field.

12.2.2 The Time-Averaged Geomagnetic Field

A fundamental assumption of paleomagnetism is that the time-averaged geomagnetic field corresponds to that of an geocentric axial dipole. The data in support of this important hypothesis come partly from studies of secular variation and partly from paleomagnetic observations in young rocks and sediments.

The dipole and non-dipole parts of the historic geomagnetic field are known to change slowly with time. Spherical harmonic analysis has shown that the axis of the geocentric inclined dipole has drifted slowly westward at about 0.044–0.14 deg yr^{-1} in the last 400 year (Fig. 11.13). This would correspond to a complete circuit about the rotation axis in 2500–8000 years. The rates of westward drift of the historic non-dipole field are around 0.22–0.66 deg yr^{-1}, giving a periodicity of 550–1650 years. However, it is important to keep in mind that the historic records of the secular variation of the geomagnetic field cover only a fragment of a complete circuit, which is not enough to confirm cyclicity or to estimate a period. The record of earlier magnetic field intensity and direction must be inferred from *archeomagnetism.*

12.2.2.1 Archeomagnetic Records of Secular Variation

Archeomagnetism is the study of the geomagnetic field recorded in dateable historical artifacts. The age of an archeological relict, such as a pot or vase, can often be determined with reliable precision. The pot, and the oven in which it was fired, may carry a TRM acquired during cooling. The intensity and direction of the field recorded by the artifacts provides an archeomagnetic record of secular variation.

A magnetic technique devised by Thellier in 1937 is used to find the intensity of the magnetic field in which an object acquired a TRM. Assuming a dipole field, the strength of the dipole magnetic moment can be inferred. When applied to rocks or ancient artifacts, the analysis delivers a *paleointensity* determination. The variation in strength of the geomagnetic field during the past 7000 years is shown by the results of 3188 paleointensity measurements on dated archeological and geological samples (Fig. 12.12). The number of archeological relicts decreases as one goes back in time, so to obtain significant mean values the data are grouped in 1000-year intervals before 2000 BC and in 500-year intervals after 2000 BC. Unfortunately, these time intervals are too short for the geomagnetic field to average to a dipole field.

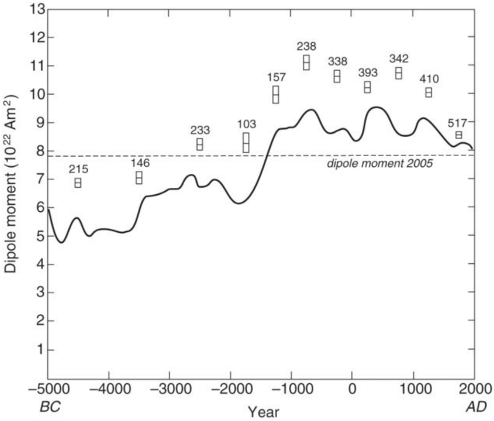


Fig. 12.12 Secular variation of the virtual geocentric dipole moment. The mean paleointensity of each time interval is plotted at the mid-point of each box, whose height represents the 95% confidence limits of the mean. The numbers indicate how many data were averaged in each time interval. The continuous curve is obtained from a global spherical harmonic analysis and the results are smoothed with a spline function.

Source: M. Korte, and C.G. Constable, The geomagnetic dipole moment over the last 7000 years – new results from a global model, *Earth Planet. Sci. Lett.*, 236, 348–358, 2005.

The total field at the time of cooling, recorded by archeological relicts, contains a non-dipole component. However, by combining the archeological paleointensity data with 13,080 inclination and 16,085 declination data from lake sediments, a global field model was calculated with Schmidt coefficients (Box 11.3) up to degree and order 10. The coefficients with $n = 1$ from this model describe the dipole magnetic moment. A smooth continuous record of the variation of the strength of the geomagnetic dipole during the past 7000 years is observed (Fig. 12.12). This record is displaced to slightly lower values than the direct measurements of paleointensity (boxes in the figure), because it is free of non-dipole components. The dipole moment has fluctuated in the past 7000 years, but the record is too short to establish if the changes are cyclical.

If the attitude of a pot during firing is known or can be assumed, the inclination of the ancient magnetic field in which the artifact was made can be deduced. The same considerations apply to lava flows that can be dated from historic records. The secular record of paleoinclination during the past 2000 years is available for two regions in which extensive archeomagnetic studies have been carried out: southeastern Europe and southwestern Japan. These regions are 110° apart in longitude but lie in similar latitude ranges, 35–40°N. Analysis of paleoinclinations suggests that the extreme values appear to occur about 400 years earlier in Japan than in Europe. The equivalent period for a full circuit of the globe is 1300 years. The observations are interpreted as the result of *westward drift* of foci of the non-dipole field past the sampling site.

The dipole field also displays westward drift, which implies systematic changes in the equatorial components of the dipole. If the changes are approximately cyclical, the mean long-term strength of the equatorial dipole measured at any given site would average to zero within a few cycles, i.e., several tens of thousands of years. Similarly, if the secular variation of the non-dipole field can be assumed to be roughly cyclical, its mean value should average to zero within a few thousand years. An averaging interval of around 10^4 years should eliminate cyclical changes, leaving the axial dipole as the only persistent long-term component of the geomagnetic field. The long-term equivalence of the Earth’s magnetic field with that of a dipole located at the center of the Earth and oriented along the rotation axis is a fundamental tenet of paleomagnetism; it is called the *geocentric axial dipole hypothesis*.

12.2.2.2 The Geocentric Axial Dipole Hypothesis

Paleomagnetic results from modern deep-sea sediments and young igneous and sedimentary rocks provide the strongest evidence in support of the geocentric axial dipole (GAD) hypothesis. Pelagic sediments are deposited extremely slowly in the deep ocean basins. The sediments acquire a post-depositional remanent magnetization (pDRM), which is an accurate record of the depositional field direction. The slow deposition ensures thorough averaging of the magnetic field. For example, at pelagic sedimentation rates a typical one-inch thick sample of deep-sea sediment averages paleomagnetic directions acquired during 2500–25,000 years of deposition. To test the GAD hypothesis the inclination observed in modern deep-sea sediment samples were compared with the inclination expected for the latitude of the sampling site. The relationship between dipole field inclination I, magnetic colatitude θ, and latitude λ is illustrated graphically in Fig. 12.13a. The mean inclinations of remanent magnetization in 186 deep-sea sediment cores of Plio-Pleistocene age taken from sites at different latitudes in the northern and southern hemispheres agree well with the values predicted by the theoretical curve for the GAD hypothesis (Fig. 12.13b).

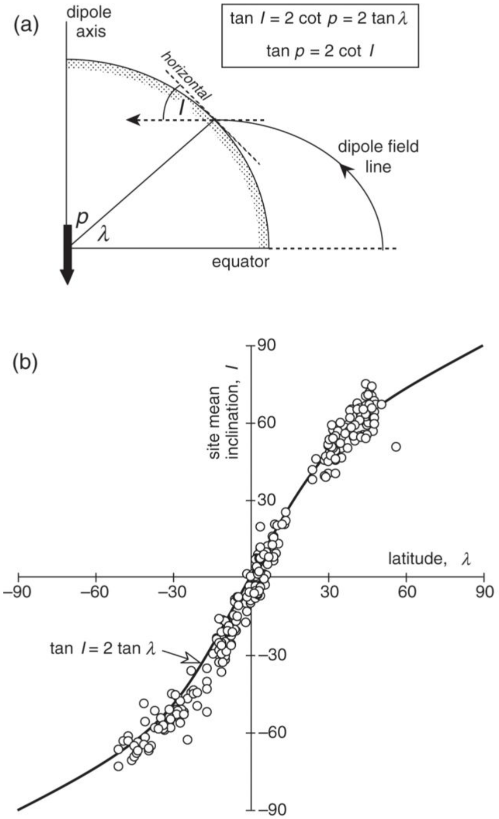


Fig. 12.13 (a) The geocentric axial dipole hypothesis predicts the relationship tan $I = 2$ tan λ between the inclination I of a dipole field and the magnetic latitude λ. (b) The inclinations measured in 186 modern deep-sea sediment cores agree well with the theoretical curve.

Data source: D.A. Schneider and D.V. Kent, The time-averaged paleomagnetic field, *Rev. Geophysics*, 28, 71–96, 1990.

Paleomagnetic studies in some terrestrial rocks present a slight complication to the GAD hypothesis. The calculated locations of paleomagnetic poles (Section 12.2.4.1) for any given site may be slightly displaced by a few degrees beyond the rotation axis. This could result if the measured inclination was flatter than it should be for a GAD field. The inclination flattening by a few degrees could be caused by compaction of sediments or by magnetic anisotropy in lavas, each of which reduces the inclination sufficiently to account for the observed effect. The flattening causes the distribution of poles to be elongated. This characteristic has been exploited to correct paleomagnetic data that have experienced inclination flattening.

According to the GAD hypothesis, the rock magnetizations in a paleomagnetic study average the non-dipole components of the geomagnetic field to zero. However, it is possible that the averaging of the field is inadequate in some cases and that axially symmetric, zonal components of the field are persistent (Box 11.3). The next strongest terms (the axial quadrupole, g_2^0 , and octupole, g_3^0) are the most likely to affect calculations of paleomagnetic pole positions. The octupole term in particular has the effect of causing shallow inclinations. It can be corrected by the same technique used to correct for inclination flattening in sediments or anisotropic lavas.

12.2.3 Methods of Paleomagnetism

Paleomagnetic analysis is designed to fulfill the requirement that the mean paleomagnetic pole position derived for a collection of rocks should represent the geocentric axial dipole. A rock formation is sampled on a hierarchical scheme designed to eliminate or minimize non-systematic errors and to average out the effects of secular variation of the paleomagnetic field. At each hierarchical level, averaging and statistical analysis are carried out on the remanent magnetization vectors. Ideally, a paleomagnetic collection should contain a large number of samples per site, but in practice about 6–10 samples are usually enough. The mean values of typically 10–20 sites from the same formation are averaged to get a mean paleomagnetic direction for a formation or region.

Paleomagnetism requires that the natural remanent magnetization (NRM) of a rock was acquired at the time of formation of the rock (or at a known time in its history), and has since remained unaltered. In fact, the NRM is usually made up of several components acquired at different times, including during the procedures of sampling and preparation. Laboratory techniques must be applied that eliminate the undesirable components and isolate the primary magnetization. This process is loosely called “magnetic cleaning.”

The presentation of paleomagnetic directions measured in rock samples is made with the help of stereographic projection. This is a way of plotting three-dimensional directions by projecting them onto a plane. These plots have already been encountered in the analysis of first-motion studies of earthquakes (Section 7.2.5.2). A direction is identified by the point where it intersects a unit sphere centered at the observation site. This converts a set of directions to a set of points on the surface of a sphere. The intersection point is then projected onto the horizontal plane to plot a stereogram. Directions with positive (downward) inclinations are plotted as lower hemisphere projections; directions with negative (upward) inclinations are plotted with a different symbol as upper hemisphere projections.

12.2.3.1 Measurement of Remanent Magnetization

The natural remanent magnetization of a rock is measured under carefully controlled laboratory conditions with a sensitive magnetometer. Two types are common: spinner and cryogenic magnetometers.

In a **spinner magnetometer** the external magnetic field of a rock sample is detected by rotating it at ~5–10 Hz in a configuration of flux-gate sensors (Section 11.4.2.1). The rotation of the sample gives a sinusoidal output, which is digitized and stored in the memory of a small online computer. The components of magnetization in the plane normal to the rotational axis are then determined by Fourier analysis. A computer-controlled flux-gate spinner magnetometer is capable of measuring a rock magnetization of 5×10^{-5} A m^{-1} in only a few minutes.

The **cryogenic magnetometer** is the most sensitive and rapid instrument in current use. Its sensor is a coil maintained by a special refrigerator (of a type called a pulse tube cryocooler) at the temperature of liquid helium. At this temperature (4 K) the coil is superconducting. A small change of magnetic field induces a comparatively large current, which because of the superconducting condition is persistent until the sample is removed. In line with the coil is a Josephson junction, which is a quantum-mechanical device consisting of a very thin element that allows the passage of current in distinct units proportional to a quantum of magnetic flux. By counting the number of flux jumps electronically the external magnetic field of the rock specimen can be inferred, and from this its magnetization computed. Most cryogenic magnetometers contain orthogonal sets of coils and can measure two or three axes of magnetization simultaneously within a few seconds. The sensitivity of the instrument corresponds to a rock magnetization of 5×10^{-6} A m^{-1} in standard samples.

12.2.3.2 Stepwise Progressive Demagnetization

The NRM of a rock may contain several components, some related to the geological history of the rock and others to the sampling and handling procedures. It is necessary to “magnetically clean” the natural magnetization so that the structure of the NRM can be analyzed and stable components isolated. This is done in a stepwise procedure, in which progressively more and more of the original magnetization is removed. There are two main methods of doing this.

The first method is progressive *alternating field* (AF) demagnetization. An alternating magnetic field can be produced in a coil by passing an alternating current through it. The field fluctuates between equal and opposite peak values. When a rock sample is placed in the alternating magnetic field, the grain magnetic moments with coercivities less than the peak value of the field are remagnetized in a new direction, while components with higher coercivity are unaffected. The intensity of the alternating field is reduced slowly and uniformly to zero. This randomizes the part of the rock magnetization that has coercivities less than the peak value of the alternating magnetic field. The AF demagnetizing coil must be surrounded by magnetic shields or special additional coils to cancel out the Earth’s magnetic field; otherwise an anhysteretic remanent magnetization (ARM) is induced along the direction of this field.

The demagnetization procedure is repeated using successively higher values of the peak alternating field, remeasuring the remaining magnetization after each step, until the magnetization is reduced to zero. Suppose that the NRM of a sample consists of two components AB and BC with different directions and different coercivity spectra (Fig. 12.14a,b). In the early stages of progressive demagnetization (steps 1–3) the “soft” component BC is first reduced to zero. The vector measured after each step in the progressive demagnetization is the sum of the “hard” component, which has not yet been affected by the field used, and the residual part of the soft component. If the direction of the soft component is within 90° of the hard component the intensity decreases during this demagnetization interval; otherwise it may increase (Fig. 12.14c). The direction of the resultant vector changes continually in steps 1–3 (Fig. 12.14d). After removal of the soft component in step 3, higher alternating fields (steps 4–7) progressively reduce the hard component AB. During this stage the intensity decreases constantly but the direction remains consistent with little scatter, defining the “stable end-point” direction of the magnetization. The stable component of magnetization defined by magnetic cleaning is called the *characteristic* magnetization, and is usually equated with the primary component.

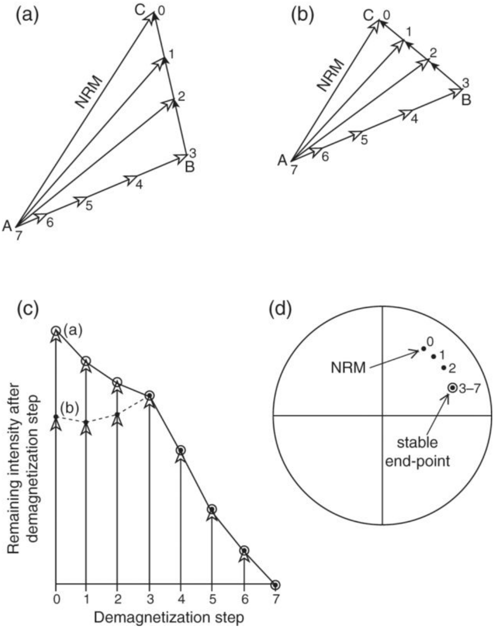

(a)
C
0
1
2
3
B
NRM
A
7
6
5
4
(b)
C
0
1
2
3
B
NRM
A
7
6
5
4
(c)
(a)
(b)
Remaining intensity after demagnetization step
0
1
2
3
4
5
6
7
Demagnetization step
(d)
NRM
0
1
2
3–7
stable end-point

Fig. 12.14 Stepwise demagnetization of a remanent magnetization consisting of two components with different ranges of stability: (a) low-stability vector BC demagnetizes before stable vector AB; (b) the same situation with a different angle between AB and BC; (c) variation of intensity for cases (a) and (b); and (d) directional changes on a stereogram. Numbers on points indicate successive demagnetization temperatures in °C.

The effectiveness of the AF demagnetization method is limited by the strongest peak field that can be produced in the demagnetizing coil. This is commonly 0.1 T, although some equipment can reach around 0.3 T. These peak fields are well below the maximum coercivity of pyrrhotite and far below the coercivities of hematite or goethite. Thus AF demagnetization is not effective in demagnetizing components carried by these minerals. The method is most commonly used for rocks that contain magnetite as the main magnetic mineral.

An alternative, and often more powerful, method of "magnetic cleaning" is by progressive *thermal* demagnetization. When a rock sample is heated to a given temperature T, thermal energy randomizes the magnetic components that have lower blocking temperatures than T. If the sample is now cooled in field-free space, this part of the NRM remains demagnetized. The heating and cooling cycle is repeated with successively higher maximum temperatures. The progressive destruction of the magnetization reveals the components present in the NRM (Fig. 12.14). This method is often more effective than AF demagnetization, because it is only necessary to heat a sample above the highest Curie temperature of its constituent minerals to destroy all of the NRM. However, if the rock contains thermally unstable magnetic minerals, irreversible changes may complicate the thermal demagnetization.

The stability of a remanent magnetization during stepwise demagnetization can be demonstrated by plotting the remaining intensity after each step against the corresponding temperature or AF field, as in Fig. 12.14c; the directional stability can be controlled simultaneously by plotting the direction after each step on a stereogram (Fig. 12.14d). The intensity plot with stereogram illustrates progressive demagnetization but it is seldom used in practice. A more rigorous analysis treats the magnetization as a vector and examines the stability of its individual components (Box 12.2). This method is an integral part of a paleomagnetic investigation.

Box 12.2 Analysis of Magnetization Components

The most powerful method of analysis of the structure and stability of a remanent magnetization is by constructing a vector diagram. The method was introduced by J. D. A. Zijderveld, a Dutch paleomagnetist, in the early 1960s. The magnetization at each stage of demagnetization is resolved into north (N), east (E), and vertical (V) components. Plots are made of the north component against the east component, and of a horizontal component (north or east) against the vertical component. This is equivalent to projecting the vector onto the horizontal plane and the north–south (or east–west) vertical plane (Fig. B12.2.1a). Components of NRM that have distinct spectra of coercivities or blocking temperatures show as linear segments on the vector demagnetization diagram.

Fig. B12.2.1b shows a Zijderveld diagram for the thermal demagnetization of a pelagic limestone. It has three distinct linear segments on the horizontal and vertical projections. The slopes of the straight lines represent NRM components with different directions. A component removed below 150°C is directed downward to the north, and so has normal polarity. The component is probably a soft overprint acquired in a recent field. A vector removed between 150°C and 300°C may represent a more ancient overprint; it is directed in a southerly, upward direction and thus has a reversed polarity. If a stable vector remains after demagnetization of less-stable fractions, it is indicated by a straight line to the origin of each half of the vector diagram. In this way the direction of the stable component of the geologically interesting remanent magnetization of the rock is determined. This is the case for the component removed between 300°C and 580°C. It is interpreted as a stable primary component acquired when the field had reversed polarity.

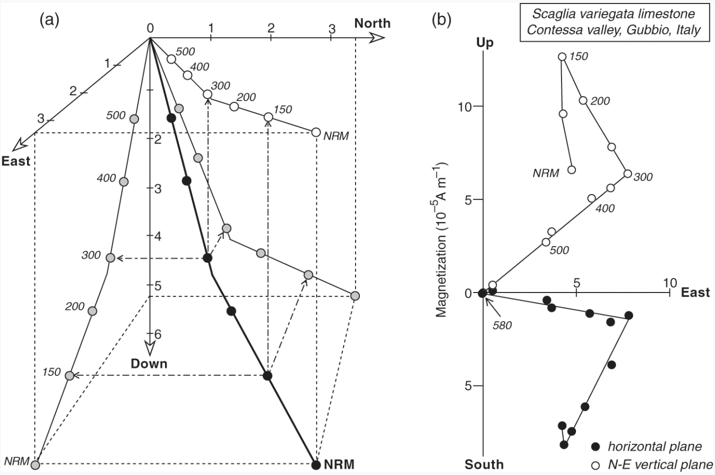


Fig. B12.2.1 The vector diagram method (Zijderveld, 1967) for analyzing progressive AF or thermal demagnetization. (a) Components of the vector remaining at each stage of demagnetization are projected as points on three orthogonal planes (horizontal, vertical north–south, and vertical east–west). (b) Vector diagram for the thermal demagnetization of a limestone sample. Magnetization components with non-overlapping spectra of thermal blocking temperatures show as linear segments.

If more than one magnetization component is present, their coercivity spectra or blocking temperature spectra may overlap partially. During demagnetization of the overlapping components the vector diagram has a curved trajectory. If the spectra overlap completely, no straight segment can be determined. In this case the sample does not have a single stable magnetization component.

12.2.3.3 Statistical Analysis of Paleomagnetic Directions

The stable paleomagnetic directions in a collection of samples are subjected to statistical analysis to determine their significance. Each direction has equal value and is regarded as a vector with unit length but with different direction from sample to sample. The end-points of the vectors lie on the surface of a unit sphere and form a distribution of points. Statistical methods for evaluating paleomagnetic directions (or the distribution of points on a sphere) were developed in 1953 by Sir Ronald Fisher, who found that the best estimate of the mean direction of a population of *N* unit vectors is their vector mean, *R*. Usually the unit vectors are not parallel. When added, the length of their resultant is shorter than if they had been parallel; the maximum length $R \leq N$. The direction of *R* is the best estimate of the mean paleomagnetic direction.

Fisher proposed that the probability density P(*θ*, κ) of the angle *θ* between an individual sample direction and the mean direction of the distribution is:

$$P(\theta, \kappa) = \frac{\kappa}{4\pi \sinh\kappa} \exp(\kappa \cos\theta). \tag{12.10}$$

The parameter κ is called the “precision parameter” or “concentration parameter.” It describes the dispersion of the directions, and is akin to the inverse of the variance of the distribution. Strictly speaking, κ is a property of an infinitely large population of directions. However, in paleomagnetic investigations only a small number of directions are usually sampled; it is assumed that they are representative for an infinite population. The parameters that are computed are approximate estimates of the true parameters of the population. Fisher showed that the best estimate (*k*) of the precision parameter κ is given by

$$k = \frac{N-1}{N-R}, \tag{12.11}$$

where *R* is the vector sum of the *N* unit vectors. When *k* (or κ) is zero, the directions are uniformly or randomly distributed. A badly scattered set of directions has a small value of *k*; large values of *k* describe tightly grouped directions.

The scatter in paleomagnetic directions is described by the angular standard deviation, which in Fisher's statistics is proportional to $1/\sqrt{k}$. However, it is usually more important to describe how well the mean direction is defined. Keep in mind that we do not know the true mean direction; we have only made an estimate of it, based on the available data. The true mean may differ by several degrees from our estimate. However, if we know that with 95% certainty the true mean lies within, say, 7°of our estimate, we can draw a cone with a semi-angle of 7° about the estimated mean direction. The cone is said to define the confidence limits of the mean at the 95% probability level. The size of the confidence limit depends on the number of directions *N* in the distribution and their dispersion parameter *k*. The semi-angle of the cone of confidence is denoted *α*95 and is given approximately by

$$\alpha_{95} = \frac{140}{\sqrt{Nk}}.$$

(12.12)

We could select any level of confidence to describe how well the mean is defined. However, two levels are common in statistics; the 95% (significant) and the 99% (highly significant) levels. In paleomagnetism the level of 95% confidence is used. This means that there is a 95% probability that the true mean of the distribution lies within this cone about the estimated mean direction.

12.2.3.4 Field Tests of Magnetization Stability

If possible, paleomagnetic sampling includes a field test that can establish the stability of the magnetization over geological time. This was especially important in the early days of paleomagnetism, when the laboratory techniques of “magnetic cleaning” were not yet available. Pioneering researchers devised some ingenious tests of paleomagnetic stability based on field observations. The fold test and reversals test still serve as the best ways to demonstrate the stability of a remanent magnetization through the eons of geological time and to verify the timing of its acquisition.

The **fold test** is perhaps the most important paleomagnetic field test. It is applied to samples taken from beds that were originally horizontal and have been tilted by later tectonic effects. If the paleomagnetic direction in the rock is stable, it will experience the same rigid body rotation as the tilted strata; its direction will vary around the fold (Fig. 12.15a, layer A). This is called a pre-folding magnetization. On the other hand, if the magnetization was acquired by the rock after it was folded, it will have a uniform direction at all points of the fold (Fig. 12.15a, layer B). This is called a post-folding magnetization. A third situation is common, in which the magnetization is acquired during the tectonic event; in this case the direction of magnetization changes around the fold but by a smaller amount than the folding. This is called a synfolding magnetization.

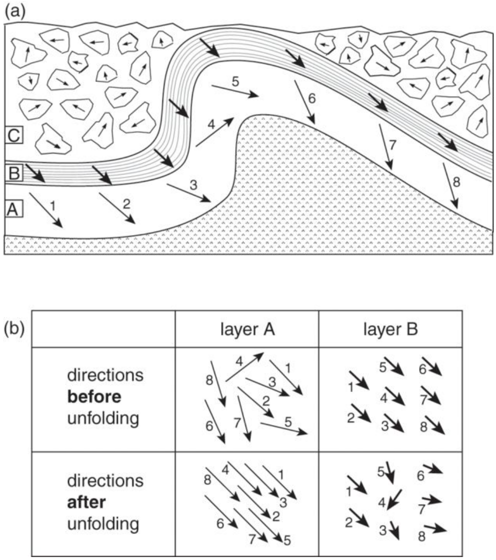

(a)
C
B
A
1
2
3
4
5
6
7
8
(b)
layer A
layer B
directions **before** unfolding
directions **after** unfolding

Fig. 12.15 (a) Magnetization directions (arrows) around a fold in stable (A) and unstable layers (B), and in stably magnetized conglomerate cobbles (C). (b) Comparison of directions in the stable layer (A) and unstable layer (B) before and after unfolding.

In practice, the fold test consists of comparing the directions before applying tilt corrections with the directions after unfolding the tilted beds. If samples with a stable magnetization are taken from all parts of the fold, their uncorrected directions are smeared out. After correction for bedding tilt, the dispersion of directions should be reduced, and the directions should group around a common direction, which is the pre-folding direction of the formation (Fig. 12.15b, layer A). This is called a positive fold test. If the magnetization is unstable or is due to post-folding re-magnetization, the tilt corrections will increase the scatter of the distribution of directions (Fig. 12.15b, layer B); this is a negative fold test.

The result of a fold test is shown in Fig. 12.16 for 12 sites of the Scaglia Rossa limestone from the central Apennines in southern Umbria (Italy). The sites were collected on different limbs of long anticlinal structures. Mean directions were computed for about 10–12 "magnetically cleaned" samples at each site. The uncorrected directions are quite scattered, with a confidence cone ($\alpha 95$) equal to 11°. After correcting each site for the local tilt of the bedding, the data are much better grouped, and the confidence cone is reduced to 6°. The increase in the concentration parameter from 14.6 to 46.3 is significant statistically, indicating that the magnetization was acquired before the folding.

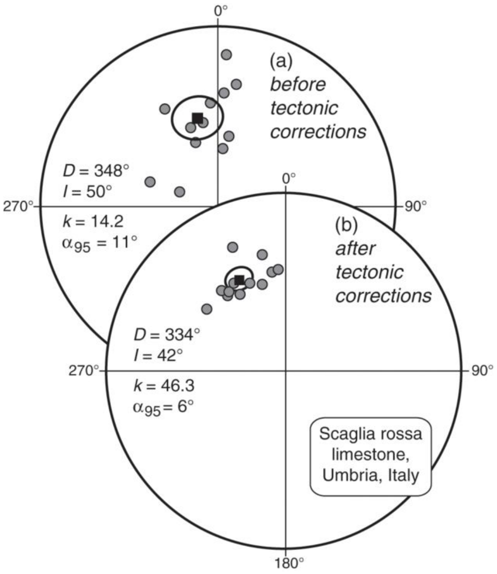


Fig. 12.16 Example of a positive fold test in 12 sites of the Scaglia Rossa limestone from southern Umbria, Italy. The directions (a) before correcting for local bedding tilt are more scattered than (b) the tilt-corrected directions.

The **conglomerate test** is a field test of stability that is rather seldom used. Suppose that we are investigating a limestone formation and that we discover a conglomerate containing cobbles of the limestone (Fig. 12.15a, layer C). Assuming that the cobbles have been randomly reoriented by the processes of erosion, transport, and re-deposition, their paleomagnetic directions, if stable, should be randomly distributed. If a systematic direction is found, the magnetization of the limestone may have a large secondary component.

The **reversals test** can be applied when the paleomagnetic samples represent a large enough time interval (>10 ka) to have recorded normal and reversed polarities of the magnetic field. Remanent magnetizations acquired within successive intervals of constant polarity of the Earth’s magnetic field should be exactly antiparallel. Let the normal magnetization be represented by the vector **N** and the reversed magnetization by the vector **R** (Fig.

12.17a). The presence of an unremoved secondary component, represented by the vector **S**, will give resultant normal and reversed directions that are no longer antiparallel. If it is possible to clean the directions magnetically, the antipodal normal and reversed directions (**N** and **R**) should be recovered. If magnetic cleaning is inadequate, a residual part of the unremoved secondary component may spoil the antiparallelism.

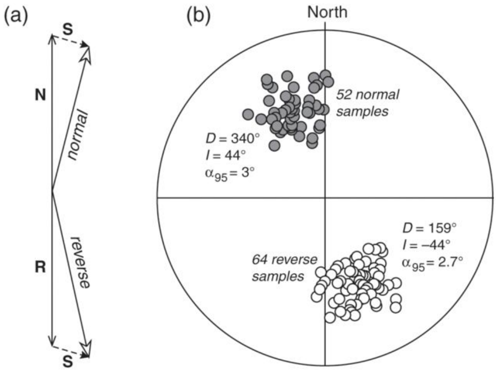


Fig. 12.17 The reversals test. (a) The presence of a secondary component **S** can cause the directions of normal and reversely magnetized samples to differ from the antiparallel **N** and **R** directions. (b) A positive fold test for samples of the *scaglia cinerea* marly limestone in a site near Gubbio, Italy. The mean directions of the normal and reverse samples are almost exactly opposite each other.

An example in which the reversals test shows successful "cleaning" is shown in Fig. 12.17b for samples from a single site in the *scaglia cinerea* formation, a marly limestone of Oligocene age in central Italy. The vector mean directions of 52 normal samples has $D_N = 340°$, $I_N = 44°$, and $\alpha_{95} = 3°$; the mean of 64 reversed polarity samples has $D_R = 159°$, $I_R = -44°$, and $\alpha_{95} = 2.7°$. These mean directions differ by almost exactly 180°, i.e., the sets of normal and reversed polarity directions are almost exactly opposite each other. The samples contain no significant contamination from secondary magnetizations that prevent them from being antipodal.

12.2.4 Paleomagnetism and Tectonics

When the site-mean directions from several sites of the Maiolica limestone are compared throughout a large region of the Umbrian Apennine mountain belt, the *declinations* are found to be more strongly dispersed than the inclinations. The smeared declinations reflect small rotations of each site about a vertical axis, the result of regional tectonism in the area of investigation. This illustrates an important application of paleomagnetism; the description of tectonic rotations that would otherwise be difficult or impossible to observe in the field. In addition to interpreting local and regional tectonics, paleomagnetism allows the analysis of past motions of continental regions. To make use of paleomagnetic data for analyzing such large-scale tectonics the observed directions of magnetization must be compared to suitable reference directions. These can be computed for a particular location if it is known where the relevant paleomagnetic pole was located in the geological past.

The effects of large-scale tectonics on paleomagnetic directions is illustrated by comparing paleomagnetic data of Late Paleozoic to Late Cretaceous age from central Europe with similar data from southern Europe (Fig. 12.18). During this long time interval the European and African plate both experienced large displacements. As a result, the inclinations measured at the indicated sites show large variations. However, the reference declinations for sites in Europe do not vary much during this time; this is also true for the African reference declinations. There is a large difference between the north–northeast pointing European reference declinations and the northwest-directed African reference declinations. North and west of a crude line through the alpine chain the observed paleomagnetic declinations agree with those expected for the European continent. However, the declinations observed south of the Alps, on the Italian peninsula and in Sicily are oriented toward the northwest, in agreement with directions expected for the *African* continent. The pattern of paleomagnetic directions support a tectonic interpretation of the Italian peninsula and adjacent regions of the Adriatic as a northern promontory of the African plate. Although the differences in the declination pattern are striking, there is a certain amount of leeway in the interpretation. This is mainly because the reference directions of Africa are derived from paleomagnetic pole locations that are not very well defined for some time intervals.

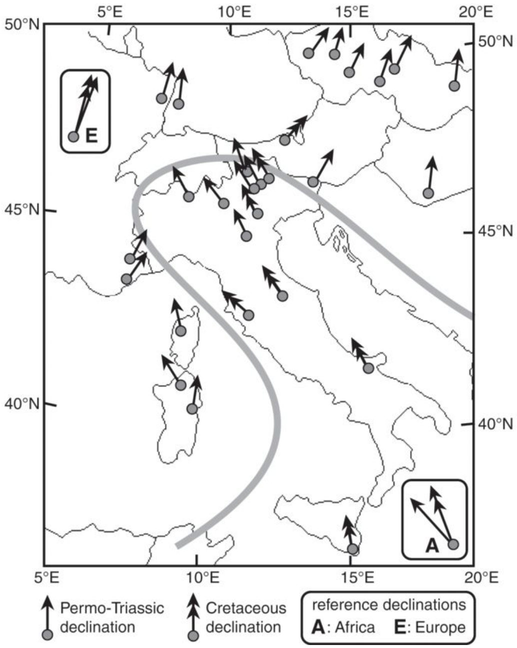

5°E
10°E
15°E
20°E
50°N
45°N
40°N
E
A
Permo-Triassic declination
Cretaceous declination
reference declinations
A: Africa
E: Europe

Fig. 12.18 Declinations of Permo-Triassic and Cretaceous rocks from Italy differ systematically from those of Europe north and west of the Alps, but agree well with directions predicted for the African plate. A hypothetical outline of the Adriatic promontory to the African plate is suggested by the shaded line.

12.2.4.1 Location of the Virtual Geomagnetic Pole

Paleomagnetic analysis assumes the validity of the geocentric axial dipole (GAD) hypothesis (Section 12.2.2.2). With this assumption it is possible to calculate where the geomagnetic pole would need to be in order to produce the observed declination *D* and inclination *I* of the remanent magnetization measured in a rock. This hypothetical location is called the *virtual geomagnetic pole* (VGP) position.

The method of computation of the VGP position is illustrated in Fig. 12.19. First, from the inclination of the magnetization (i.e., the paleofield), we can calculate the angular distance p of the site from the VGP at the time the rock magnetization was acquired, assuming a GAD field. The relationship between inclination I and polar angle p is given in Eq. (11.12) and illustrated in Fig. 12.13a.

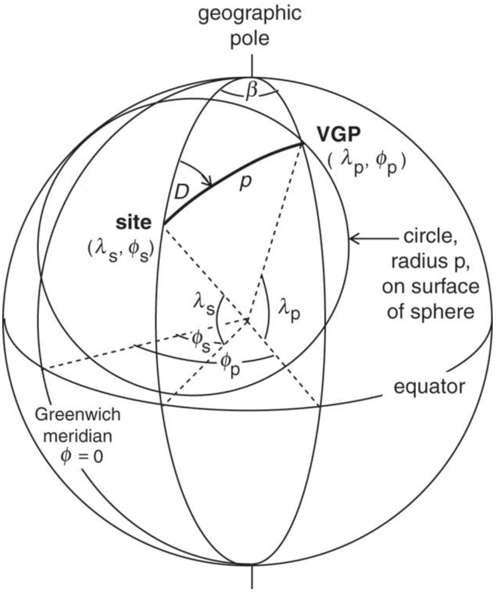

geographic pole
β
VGP
(λ_p, ϕ_p)
D
p
site
(λ_s, ϕ_s)
circle, radius p, on surface of sphere
λ_s
λ_p
ϕ_s
ϕ_p
equator
Greenwich meridian
$\phi = 0$

Fig. 12.19 The method of locating the virtual geomagnetic pole (VGP) from the declination D and inclination I measured at a site.

Source: T. Nagata, *Rock Magnetism*, 350 pp., Tokyo: Maruzen, 1961.

The distance p is the radius of a small circle centered on the paleomagnetic sampling site, say at latitude λ_s and longitude ϕ_s. The circle is the locus of all possible VGP positions that could give the observed inclination I at the site. We next have to decide which point on the small circle is the VGP position. This is decided by the declination of the remanent magnetization, which represents the angle between geographic north and the horizontal direction to the ancient magnetic pole. The declination defines a meridian (or great circle) which passes through the sampling site and makes an angle *D* with the north–south meridian (Fig. 12.19). The place where this great circle intersects the small circle around the sampling site is the location of the virtual geomagnetic pole. Its latitude (λ_p) and longitude (ϕ_p) in present-day geographic coordinates are computed from trigonometric formulae (Box 12.3):

$$\sin\lambda_p = \sin\lambda_s \cos p + \cos\lambda_s \sin p \cos D, \tag{12.13}$$

$$\begin{aligned} \phi_p &= \phi_s + \beta, && \text{for} \quad \cos p \geq \sin\lambda_s \sin\lambda_p \\ \phi_p &= \phi_s + 180 - \beta, && \text{for} \quad \cos p < \sin\lambda_s \sin\lambda_p, \end{aligned} \tag{12.14}$$

where

$$\sin\beta = \frac{\sin p \sin D}{\cos\lambda_p}. \tag{12.15}$$

Box 12.3 Virtual Geomagnetic Pole Location

The sine and cosine relationships between the sides and angles of spherical triangles (Box 2.1) may be applied to the spherical triangle in Fig. 12.19 – whose corners correspond to the sampling site, VGP position, and geographic pole – to determine the unknown latitude and longitude of the virtual paleomagnetic pole. The length of the side between the site and the geographic pole, measured in degrees of arc, is (90 – λ_s), where λ_s is the site latitude. The length of the side between the VGP position and the geographic pole is (90 – λ_p), where λ_p is the VGP latitude. The length of the third side of the triangle is *p*. The direction to the VGP position from the site is the declination D, which is opposite the side of length $(90 - \lambda_p)$. These values are substituted in the law of cosines (Box 2.1, Eq. 4) as follows:

$$\begin{aligned}
&a = (90 - \lambda_p);\ \cos a = \sin\lambda_p \\
&b = (90 - \lambda_s);\ \cos b = \sin\lambda_s;\ \sin b = \cos\lambda_s \\
&c = p;\ A = D,
\end{aligned} \tag{1}$$

$$\sin\lambda_p = \sin\lambda_s\ \cos p + \cos\lambda_s\ \sin p\ \cos D. \tag{2}$$

The great circles through the site and the VGP location meet at the geographic pole where they form an angle, for which there are two possibilities: the acute angle β, or its obtuse equivalent $(180 - \beta)$. As a result, the solution for the longitude ϕ_p of the paleomagnetic pole involves two steps.

First, applying the law of sines (Box 2.1, Eq. 3) to the spherical triangle gives

$$\frac{\sin\beta}{\sin p} = \frac{\sin D}{\sin(90 - \lambda_p)}, \tag{3}$$

from which the magnitude of the angle β may be obtained:

$$\sin\beta = \frac{\sin p\ \sin D}{\cos\lambda_p}. \tag{4}$$

Next, in order to decide whether the solution requires β or

(180 – *β*), the law of cosines is again applied to the spherical triangle, this time for side *p*, with the following substitutions:

$$\begin{aligned} &a = p; A = \beta \\ &b = (90 - \lambda_s); \cos b = \sin\lambda_s; \sin b = \cos\lambda_s \\ &c = (90 - \lambda_p); \cos c = \sin\lambda_p; \sin c = \cos\lambda_p. \end{aligned} \tag{5}$$

These give

$$\cos p = \sin\lambda_s \sin\lambda_p + \cos\lambda_s \cos\lambda_p \cos\beta, \tag{6}$$

$$(\cos p - \sin\lambda_s \sin\lambda_p) = \cos\lambda_s \cos\lambda_p \cos\beta. \tag{7}$$

On the right-hand side of Eq. (7), $\cos\lambda_s$ and $\cos\lambda_p$ are always positive, so $\cos\beta$ must take the sign of the left-hand side of the equation. This gives the two possibilities for the value of $(\phi_p - \phi_s)$, namely, *β* or (180 – *β*. Thus,

$$\begin{aligned} &(\phi_p - \phi_s) = \beta, \text{ for } (\cos p - \sin\lambda_s \sin\lambda_p) \geq 0 \\ &(\phi_p - \phi_s) = 180 - \beta, \text{ for } (\cos p - \sin\lambda_s \sin\lambda_p) < 0. \end{aligned} \tag{8}$$

Assuming the GAD hypothesis, all locations on the same latitude (i.e., at the same distance *p* from the rotation axis) acquire magnetizations with the same inclination and with zero declination, because the axial dipole field lines through the site lead to the rotation axis. The longitude of the site (its position on the circle of latitude) cannot be determined. The declination measured later at a given site is the expression of any change of azimuthal orientation resulting from local tectonic motion or from large-scale continental displacement.

Virtual geomagnetic pole positions are used to compute a *paleomagnetic pole* by statistically averaging data from a number of samples or sites. For example, when a lava flow extrudes, it cools rapidly in geologic terms. Each sample from the lava becomes magnetized in a short interval of time, and thus the recorded field direction is that of the total geomagnetic field at the site at the time of the flow, and includes the axial dipole, non-axial dipole, and non-dipole components. As a result, the computed VGP may not coincide with the rotation axis. However, if data are collected from several flows of different ages, each carries a slightly different record of the field. The computed VGP positions are different from flow to flow, giving a scattered distribution. If a sufficiently large number of lava flows of the same age are sampled, covering a long enough period of time to average the non-dipole and non-axial dipole components to zero, the mean direction of the collection then corresponds to a GAD field. The pole position calculated from the mean direction of a collection of flows represents an averaged value of the field, and is called a *paleomagnetic pole*.

The VGP represents a spot estimate of the field, including non-axial dipole components; the paleomagnetic pole represents an averaged field. The GAD hypothesis maintains that, if data are averaged over a long enough interval of time, the mean paleomagnetic pole position will coincide with the Earth's rotation axis.

12.2.4.2 Apparent Polar Wander Paths

The paleomagnetic pole positions determined in studies of volcanic and sedimentary rocks in North America of Neogene age (i.e., younger than 20 Ma) support the GAD hypothesis (Fig. 12.20). The young paleomagnetic poles are clustered around the geographic pole and not around the present-day geomagnetic pole. Statistical analysis shows that the mean of the young paleomagnetic poles does not differ significantly from the geographic pole. However, when paleomagnetic pole positions are calculated for old rocks from the same continent, they group far away from the geographic pole. The Permian paleomagnetic poles from North America are located in northeastern Asia, about 45° away from the geographic pole. If the axial dipole hypothesis is valid for rocks of all ages, the pole distributions imply that the geographic pole for North America in the Permian period (about 250–300 Myr ago) lay far from its present position. An alternative interpretation is that the geographic pole has not changed, but North America has moved relative to the pole; the position about which the Permian poles now cluster was on the rotation axis in the Permian period. The North American continent has subsequently moved to its present-day position with regard to the rotation axis.

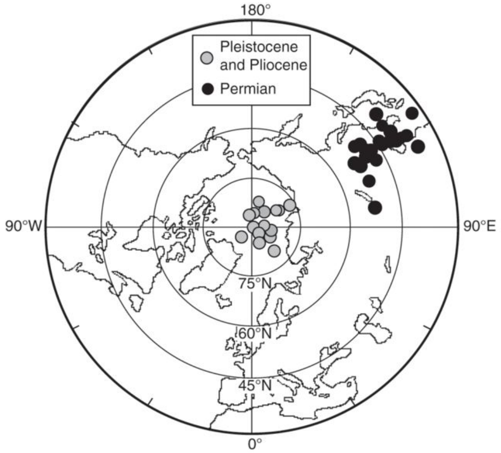


Fig. 12.20 Locations of North American paleomagnetic poles. Pliocene and Quaternary poles lie close to the present-day geographic pole, while Permian poles are located at about 45°N in northeastern Asia.

Data source: T.H.Torsvik et al., Phanerozoic polar wander, palaeogeography, and dynamics, *Earth-Science Rev.*, 114, 325–368, 2012.

Paleomagnetic data allow us to resolve the ambiguity. If paleomagnetic pole positions are computed for rocks of different ages from the same continent, they plot systematically along an irregular, curved path. It appears as though the paleomagnetic pole has moved slowly along this path toward the present rotation axis. The apparent motion of the paleomagnetic pole is called *apparent polar wander* (APW) and the path is called an *apparent polar wander path*. First introduced by K. M. Creer in 1954, when paleomagnetism was in its infancy, the APW is an important tool for paleogeographic reconstructions. The paleomagnetic data from a particular continent define a unique APW path for that continent, and each continent has a different APW path. Thus, we have a European APW path, African APW path, North American APW path, and so on.

The APW paths for Europe and North America from the Ordovician to the Middle Jurassic are clearly distinct curves (Fig. 12.21a). Keeping the geocentric axial dipole hypothesis in mind, it is obviously impossible that the paleomagnetic pole (i.e., the Earth's rotation axis) could have moved simultaneously along two different APW paths. The two APW paths evidently represent the separate motions of the European and North American continents relative to the rotation axis. They constitute paleomagnetic evidence for "continental drift."

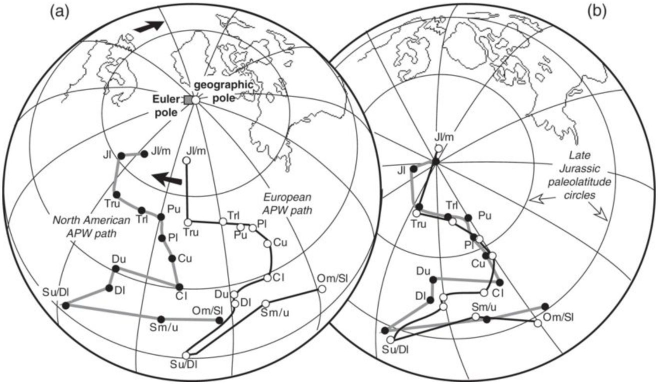

(a)
geographic pole
Euler pole
Jl
Jl/m
Jl/m
Tru
Trl
Pu
North American APW path
Pl
Tru
Trl
Pu
Pl
European APW path
Cu
Cu
Du
Cl
Su/Dl
Dl
Cl
Du
Dl
Om/Sl
Om/Sl
Sm/u
Sm/u
Su/Dl
(b)
Jl/m
Jl
Late Jurassic paleolatitude circles
Trl
Pu
Tru
Pl
Cu
Du
Dl
Cl
Sm/u
Om/Sl
Su/Dl

Fig. 12.21 (a) The Ordovician to Jurassic segments of the North American and European APW paths. (b) The same APW paths after rotating Europe by 38° clockwise about the Euler rotation pole at 88.5°N, 27.7°E, marked by the square symbol in (a).

Source: R. Van der Voo, Phanerozoic paleomagnetic poles from Europe and North America and comparisons with continental reconstructions, *Rev. Geophys.* 28, 167–206, 1990.

12.2.4.3 Paleogeographic Reconstructions Using APW Paths

The European and North American APW paths have very similar shapes for the time before the Late Jurassic, particularly from the Upper Carboniferous to the Upper Triassic. It is possible to overlay these two segments of the APW paths by moving Europe (including Russia west of the Ural mountains) and North America into different positions relative to each other (Fig. 12.21b). For the time represented by the overlap of the APW paths the two continents formed part of a larger "supercontinent," called *Laurasia* (also referred to as *Laurussia* or *Euramerica* in some reconstructions). The present separation of the APW paths (Fig. 12.21a) is evidence for relative plate tectonic motion between Europe and North America since the end of the interval for which the APW paths overlap well, i.e., since the Early Jurassic.

In order to bring the two APW paths into coincidence we have to move Europe relative to North America (or vice versa) so as to close the present gap between the continents. As shown in Section 2.9, the relative motion of plates on the surface of the spherical Earth is equivalent to a relative rotation about an *Euler pole of rotation*. The computer-generated "Bullard" fit of the 500 fathom contour lines on opposites sides of the North Atlantic Ocean (Section 2.2.2) can be obtained by displacing Europe toward North America by a clockwise rotation through 38° about the Euler pole located at 88.5°N, 27.7°E, which by coincidence is very close to the present-day geographical pole (Fig. 12.21a). The APW path of a continent is constrained to move with the continent. If the European APW path is also rotated by 38° clockwise about the same Euler pole, the observed overlap of the Upper Carboniferous to Upper Triassic sections of the European and North American APW paths is obtained. Later segments of the paths diverge, indicating relative motion between the continents. The Late Jurassic pole position corresponds to the Earth's rotation axis at that time. Circles of paleolatitude about the North American paleomagnetic pole emphasize the paleogeographic reconstruction of the relative positions of Europe and America in the Late Jurassic.

The interpretation of APW paths is not always as clear-cut as in this example. Consider what happens when the paleomagnetic pole P for a continental plate C is rotated through an angle Ω about an Euler pole of rotation, E. First, suppose that the plate lies between the paleomagnetic pole and the rotation pole (Fig. 12.22a). The rotation of the plate from C to C′ causes the paleomagnetic pole P to move to P ′. The arc PP′ of the polar motion is longer than the arc CC′ of the true plate motion. Next, consider what happens when the paleomagnetic pole lies between the plate and the Euler pole (Fig. 12.22b). In this case the plate moves through a large distance, but the paleomagnetic pole moves only a small distance. In the extreme case where the paleomagnetic and Euler poles coincide, the plate rotation does not displace the paleomagnetic pole at all. Under these special conditions plate motion leaves no trace in the APW path of the plate.

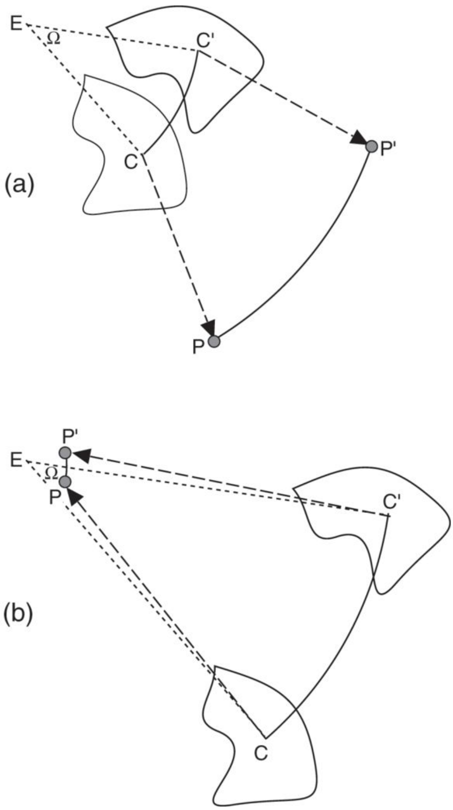


Fig. 12.22 Rotation of a continental plate C about an Euler pole E

displaces the paleomagnetic pole P (a) by a large amount if P is further from E than the continent C, and (b) by only a small amount when P lies close to E.

Clearly, the interpretation of an APW path as a record of plate motion relative to the geographic axis must be made with caution. The rate of motion of the pole along an APW path cannot be simply equated with the rate of motion of the parent continent or global plate. It follows that similarity of APW paths does not imply a unique solution for former relative plate positions. However, if two continents once belonged to the same plate for some length of time, they should have acquired the same APW path for this time. Matching the present APW paths of the continents for the time they were on the same plate should give a viable reconstruction of the earlier positions of the continents relative to each other. To avoid ambiguities, additional independent evidence (such as paleoclimatic data or computer-matching of coastlines) must be utilized in conjunction with paleomagnetic data for making such reconstructions.

12.2.4.4 Paleomagnetism and Continental Drift

The nineteenth-century geologist Eduard Suess deduced the existence of a great Late Paleozoic continent, which he called Gondwana (Section 2.1). It was composed of Africa, Antarctica, Arabia, Australia, India, and South America. In 1912, Alfred Wegener went a step further by postulating that all the present continents lay close together during the Late Paleozoic, forming a single great continent that he called Pangea (Section 2.2). Wegener's concept was based on paleoclimatic evidence, the matching of Carboniferous coal belts and of regions of Paleozoic glaciation in the different continents. Subsequently, additional geological evidence for the existence of Gondwana and Pangea during the Late Paleozoic and Early Mesozoic accumulated from the fields of sedimentology, paleontology, and tectonics. The earlier great continents were presumed to have dispersed to their present-day location by the process of continental drift. Unfortunately, Wegener was unable to offer a satisfactory driving mechanism for continental drift, and some of his ideas were found to be extreme. Skepticism among geophysicists and geologists brought Wegener's theories into disrepute. With hindsight, the mechanism that produces continental drift is plate tectonics.

Interest in continental drift was re-awakened in the 1950s by the development of paleomagnetism. Soon thereafter some of the most convincing paleomagnetic evidence for continental drift was obtained from the "southern continents." Researchers found that Mesozoic paleomagnetic pole positions of the same age from these continents were very dissimilar. In landmark contributions, E. Irving showed that the paleomagnetic poles of the southern continents were incompatible with the present-day arrangement of these continents, but that they agreed much better when the continents were rearranged to conform with a Gondwana reconstruction.

Numerous paleomagnetic results from dated rocks have been collected into an extensive database that can be used to test the validity of reconstructions of former great continents at different times in their history. The reconstructions are generally not made on paleomagnetic evidence alone. Usually, a model is proposed, based on geometrical or geological grounds. The congruity of paleomagnetic pole positions from the separate continents is then evaluated in their reconstructed positions. The model is adjusted iteratively until a configuration of the continents is obtained that gives minimum scatter of the paleomagnetic poles. When possible, APW paths of the continents to be matched are determined by averaging paleomagnetic pole positions in suitable time windows, e.g., 20–40 Myr. The paths are then fitted together to produce reconstructions for the time represented by coeval segments, as in Fig. 12.21.

12.2.4.5 Absolute Plate Motions and True Polar Wander

During the displacement of a plate (i.e., when it rotates about an Euler pole), the paleomagnetic pole positions obtained from rocks on the plate describe a trajectory which is the arc of a small circle about the Euler pole (Fig. 12.23). The motion of the plate over an underlying hotspot leaves a trace that is also a small circle arc about the same hotspot. The paleomagnetic record gives the motion of plates relative to the rotation axis, whereas the hotspot record shows the plate motion over a fixed point in the mantle. If the mantle moves relative to the rotation axis, the network of hotspots – each believed to be anchored to the mantle – shifts along with it. This motion of the mantle deeper than the mobile lithosphere is called *true polar wander* (TPW). It arises when density heterogeneities in the mantle are displaced, for example as a result of isostatic readjustment accompanying removal of crustal loads when ice-sheets melt (see Fig. 5.22). The mass distribution changes and this alters the moments of inertia of the Earth, which then adjusts so as to keep the maximum moment of inertia close to the rotation axis. As a result the entire mantle may shift relative to the rotation axis, giving the appearance that the rotational pole wanders. On a longer timescale the displacement of mass during mantle convection may alter the Earth's moments of inertia (e.g., if cold lithospheric slabs are subducted into the lower mantle and form a mass anomaly).

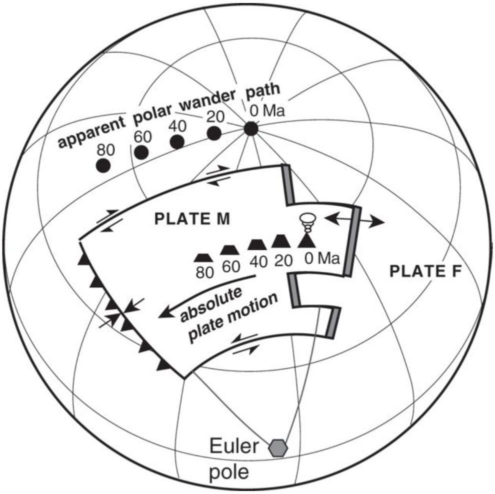


Fig. 12.23 Development of an arcuate apparent polar wander path

and hotspot trace as small circles about the same Euler pole, when a mobile plate M moves relative to a fixed plate F.

Source: R.F. Butler, *Paleomagnetism: Magnetic Domains to Geologic Terranes*, 319 pp., Boston: Blackwell Scientific Publ., 1992.

Paleomagnetism provides a means of detecting whether long-term TPW has taken place. It involves comparing paleomagnetic poles from hotspots with contemporary poles from the stable continental cratons. Consider first the possibility that TPW does not take place: Each hotspot maintains its position relative to the rotation axis. A lava that is magnetized at an active hotspot acquires a direction appropriate to the distance from the pole. If the plate moves from north to south over the stationary hotspot, a succession of islands and seamounts (A–D) is formed, which, independently of their age, have the same magnetization direction (Fig. 12.24a). Next, suppose that TPW takes place: Each hotspot moves with time relative to the rotation axis. For simplicity, let the hotspot migration also be from north to south (Fig. 12.24b). Seamount A is being formed at present and its magnetization direction corresponds to the present-day distance from the pole. However, older seamounts B, C, and D were formed closer to the pole and have progressively steeper inclinations the further south they are. The change of paleomagnetic directions with age of the volcanism along the hotspot trace is evidence for TPW.

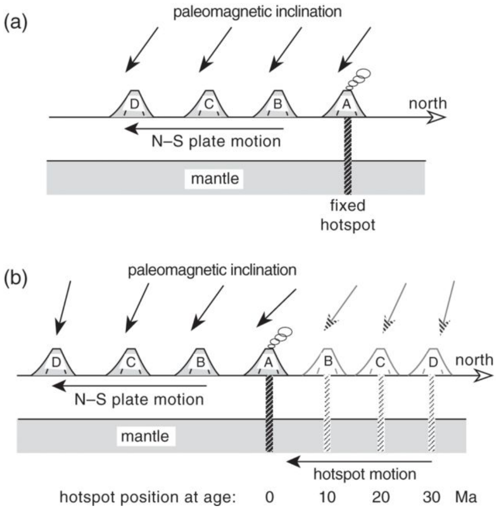

(a)
paleomagnetic inclination
D
C
B
A
north
N–S plate motion
mantle
fixed hotspot
(b)
paleomagnetic inclination
D
C
B
A
B
C
D
north
N–S plate motion
mantle
hotspot motion
hotspot position at age: 0 10 20 30 Ma

Fig. 12.24 Illustration of the effect of true polar wander on paleomagnetic inclination: (a) north–south plate motion over a stationary hotspot, (b) same plate motion over a north–south migrating hotspot. A, B, C and D are sequential positions.

To test such a hypothesis adequately, a large number of data are needed. The amount of data from a single plate, such as Africa, can be enlarged by using data from other plates. For example, in reconstructing Gondwana, South America is rotated into a matching position with Africa by a finite rotation about an Euler pole. The same rotation applied to the APW path of South America allows data from both continents to be combined. Likewise, rotations about appropriate Euler poles make the paleomagnetic records for North America and Eurasia accessible. Averaging the pooled data for age-windows 10 Myr apart gives a reconstructed paleomagnetic APW path for Africa (Fig. 12.25a). The next step is to determine the motions of plates over the network of hotspots, assuming the hotspots have not moved relative to each other. A “hotspot” APW path is obtained, which is the track of an axis in the hotspot reference frame presently at the north pole. The appearance of this track relative to Africa is shown in Fig. 12.25b.

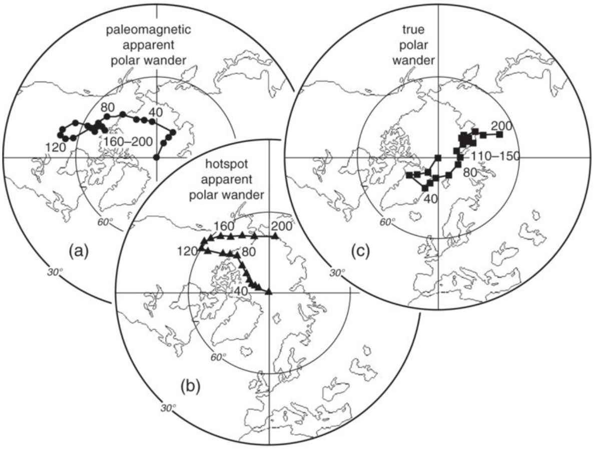

paleomagnetic apparent polar wander
80
40
120
160–200
60°
(a)
30°
hotspot apparent polar wander
160
200
120
80
40
60°
(b)
30°
true polar wander
200
110–150
80
40
60°
(c)
30°

Fig. 12.25 (a) Paleomagnetic APW path reconstructed for Africa using data from several plates. (b) Hotspot APW path (motion of an axis at the geographic pole relative to the hotspot reference frame). (c) Computed true polar wander path.

Data source: V. Courtillot and J. Besse, Magnetic field reversals, polar wander, and core-mantle coupling, *Science* 237, 1140–1147, 1987.

These calculations give records of the motion of the lithosphere relative to the pole, and of the motion of the lithosphere relative to the hotspot reference frame. The records coincide for the present time, with pole positions at the present-day rotation axis, but they diverge with age as a result of TPW. A paleomagnetic pole of a given age is now moved along a great circle (i.e., rotated about an Euler pole in the equatorial plane) until it lies on the rotation axis. If the same rotation is applied to the hotspot pole of the same age, it should fall on the rotation axis also. A discrepancy arises if the hotspot reference frame moves relative to the rotation axis. Joining locations in order of age gives a TPW path (Fig. 12.25c). The results show that TPW has indeed taken place but that its amplitude has remained less than 15°, at least for the last 150 Myr.

12.2.5 Paleozoic and Precambrian Paleomagnetism

Paleomagnetic reconstructions become more difficult and less reliable with increasing age in the geological record. The GAD hypothesis is fundamental to the credibility of paleomagnetic data and evidence indicates that the hypothesis has been valid throughout the Phanerozoic. However, this has not yet been established for most of the much longer Precambrian. Moreover, the oldest parts of the paleomagnetic record are clouded by the increasing possibility of false directions due to tectonic displacement or undetected secondary magnetization. This happens when thermal or tectonic events alter the original magnetization, so that its direction no longer corresponds to the direction acquired at the time of rock formation. Re-magnetization can affect rocks of any age, but it is recognized more readily and constitutes a less serious problem in younger rocks. With these caveats in mind some general features of paleomagnetic reconstructions of earlier continents are well established and are supported by geological and geochemical evidence.

12.2.5.1 Paleozoic Supercontinents

Paleomagnetism permits quite detailed reassembly of the supercontinents Gondwana and Laurussia in the Early Paleozoic and traces their movements before they collided in the Carboniferous to form Pangea. Geological and paleomagnetic evidence concur that, in the Cambrian period, Gondwana very likely existed as a supercontinent in essentially the configuration proposed by du Toit. It coexisted in the Early Paleozoic with three other cratonic centers: Laurentia (North America and Greenland), Baltica (northern Europe), and Siberia. Laurentia and Baltica were separated by the Iapetus Ocean, which began to close in the Ordovician (about 450 Myr ago), resulting in the collision of Laurentia and Baltica around Late Silurian time and their fusion to form the supercontinent Laurussia. The collision is expressed in the Taconic and Caledonian orogenies in North America and northern Europe. At that time the Siberian block remained a separate entity. Readjustments of the positions of the continental blocks in the Devonian produced the Acadian orogeny. The Rheic ocean which separated Gondwana and Laurussia in the Middle Devonian (about 410 Myr ago, Fig. 12.26a) closed progressively through the Carboniferous, producing the Hercynian orogeny. With the accretion of Siberia in the Late Carboniferous, the supercontinent Pangea was complete; with the exception of China and smaller blocks in eastern Asia it formed a single landmass surrounded by the Panthalassa ocean. Pangea subsequently existed throughout the Permian (Fig. 12.26b) and Triassic, until its eventual dispersal in the Early Jurassic.

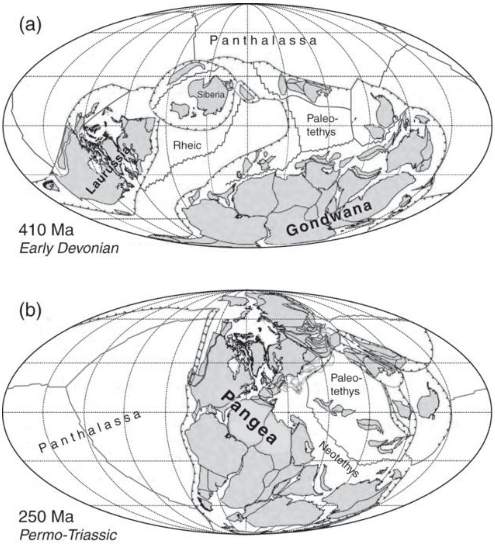

(a)
Panthalassa
Siberia
Paleo-tethys
Rheic
Laurussia
Gondwana
410 Ma
Early Devonian
(b)
Paleo-tethys
Pangea
Panthalassa
Neotethys
250 Ma
Permo-Triassic

Fig. 12.26 (a) The positions of the major continental units in the early Devonian and (b) their Permo-Triassic configuration as Pangea.

Source: Domeier, M. and Torsvik, T. H. 2014. Plate tectonics in the Late Paleozoic. *Geoscience Frontiers*, 5, 303–350.

12.2.5.2 The Break-Up of Pangea

The optimization of paleomagnetic pole positions and the analysis of APW paths is an important tool for understanding the positions of the continents, and in conjunction with independent geological data can be applied also to the past 175 Myr, to understand what happened following the break-up of Pangea. The present northern continents (North America, Europe, and adjacent Asia) formed a supercontinent known as Laurasia (similar but not identical to the Early Paleozoic Laurussia), while the southern continents (Africa, Antarctica, Australia, India, and South America) formed the supercontinent Gondwana. Between northeast Gondwana and southeast Laurasia, and open to the East, lay the *Tethys* Ocean.

The break-up of Pangea initiated the current phase of sea-floor spreading. The earliest separation began about 180 Myr ago during the Late Jurassic, when Laurasia split from Gondwana, forming the North-Central Atlantic Ocean between North America and Africa. In the Late Cretaceous North America began to separate from Europe, accompanying the opening of the Greenland sea, then, in the Cenozoic, the Norwegian Sea opened, increasing the separation of Europe from North America; this motion still continues and is observable by space geodetic methods (Fig. 2.12). Since the Cenozoic, Africa has also moved northward against Eurasia, closing the Tethys Ocean, whose western expanse became the present Mediterranean Sea.

In the Early Cretaceous the South Atlantic opened between South America and Africa, and Gondwana began to break up into its individual continents. Australia separated from Antarctica, and India moved rapidly northward at a geologically rapid rate of around 15 cm yr^{-1} to collide with, and plunge under, Eurasia about 50 Myr ago, forming the Himalayan mountain chain.

The post-Pangea plate motions can be reconstructed in remarkable detail because they took place over millions of years, during which the geomagnetic field reversed its polarity many times at irregular intervals. The mechanism of sea-floor spreading imprinted the record of alternating geomagnetic polarity in the magnetization of the oceanic crust, and the sets of lineated magnetic anomalies in different ocean basins can be analyzed to reconstruct and time the relative plate motions.

12.2.5.3 Precambrian Paleomagnetism and the Early Geodynamo

The range of time encompassed by the Precambrian – almost 4000 Myr – is about six times longer than the 541 Myr length of the Phanerozoic (see Fig. 8.2). In that time, lithosphere destruction and renewal may have taken place many times. The probability of re-magnetization events is correspondingly high, such that original magnetizations may have been overprinted by younger events. Despite the discovery of paleomagnetic directions that pass the criteria of acceptability, it has not been possible to develop APW paths. To describe relative tectonic motion between two cratonic blocks, it is necessary to compare paleomagnetic poles from rocks of the same age on each block. This requires more precise dating of the rocks than is usually possible. However, enough reliable paleomagnetic data exist to show that rotations and displacements between cratonic regions were taking place. This agrees with geochemical and geological evidence, which suggests that plate tectonic motions were perhaps active about three billion years ago. Moreover, there is strong evidence that geomagnetic reversals were taking place in the Precambrian. Russian limestones of Neoproterozoic age (1600–650 Ma) have been found to contain well-defined magnetozones with alternating polarity.

Paleomagnetic estimates of the paleointensity of the geomagnetic field have been derived from magnetite inclusions in single silicate grains that have ages of around 3.4 Ga. The Precambrian magnetic field had a weaker intensity than the present field, so the outer boundary of the magnetosphere would have been much closer to the planet. The presence of a magnetic field would have shielded the Earth against solar radiation, but less effectively than at present.

Two mechanisms drive the present-day geodynamo: thermal and compositional convection (Section 11.2.6.2). Thermal convection accompanies slow cooling of the core; compositional convection results from solidification of the inner core and the associated discharge of light elements into the core fluid. Recent calculations and experiments have shown that the thermal conductivity of iron under core conditions is much higher than previously thought. More of the core's heat could therefore be transported by conduction and the importance of thermal convection would be reduced.

The calculations also imply that the age of the inner core must be less than about 1 Ga. The existence of an older geomagnetic field with an age of at least 3.4 Ga requires an additional mechanism to drive the geodynamo. It is theorized that impacts during Earth's accretion (such as the collision with the Mars-sized body called Theia, which produced the Moon) may have allowed magnesium to be incorporated into the early core during the Earth's differentiation. Subsequent exsolution of the magnesium at the core–mantle boundary could have caused a compositional buoyancy with enough energy to power the early geodynamo.

12.3 Geomagnetic Polarity

12.3.1 Introduction

The earliest demonstration that the geomagnetic field has changed polarity in the past was made by the French scientists P. David and B. Brunhes. In 1904–1906 they described the magnetic properties of young lava flows in the Massif Central region of France. They found that clays baked by the lava flows had the same direction of remanent magnetization as the lavas. Moreover, when the magnetization direction in the lava was opposite to that of the present-day field, the same was the case in the baked clay. The opposite polarities were interpreted as evidence that the geomagnetic field can reverse its polarity.

A Japanese scientist, M. Matuyama, was the first to associate the polarity of remanent magnetization in lavas with their age, determined stratigraphically. In 1929 he reported finding young Quaternary lavas with magnetization directions close to the present-day field direction, whereas the directions in older Quaternary lavas were clustered about an antipodal direction. He also found that one of three samples of Miocene basalt was magnetized oppositely to the other two. Matuyama's interpretation was that geomagnetic polarity had changed several times during Late Tertiary time.

The idea that geomagnetic polarity could change was controversial, and for many years skeptics sought alternative interpretations. Scientists realized that the observed reversed polarities might have a mineralogical explanation. Indeed, some ferromagnetic minerals, because of their composition and structure, can acquire a TRM exactly opposite to the field direction. This mechanism is called self-reversal of magnetization. It has been described in lavas in which the ferromagnetic minerals are particular forms of titanohematite. Fortunately, it is a rather rare phenomenon. Most reversed magnetizations in rocks are records of reversals of the geomagnetic field.

To envisage what a reversal of geomagnetic polarity means, imagine that the geomagnetic dipole inverts its direction. At present the geocentric axial dipole points from the northern hemisphere toward the southern hemisphere; a polarity reversal would orient the dipole in the opposite direction. At each point on the surface the magnetic inclination I changes sign and the declination D changes by 180°; for example, a normal direction $\{I = 40^\circ,\ D = 30^\circ\}$ might change to a reverse direction $\{I = -40^\circ, D = 210^\circ\}$. A polarity reversal is a global event, experienced simultaneously all over the Earth. Thus, geomagnetic reversals provide an important method for stratigraphic correlation and dating.

12.3.1.1 Geomagnetic Polarity Transitions

The change of polarity from one sense to the opposite one is called a *polarity transition*. Paleomagnetic records of polarity transitions have been observed in radiometrically dated lava sequences and in deep-sea sediments with known deposition rates. These records indicate that the duration of a polarity transition is about 3.5–5 kyr. This is much shorter than the length of the interval of constant polarity before or after the transition, which may last for hundreds of thousands or millions of years.

The detailed behavior of the geomagnetic field during a polarity transition is not well understood. The dipole field is dominant before and after a transition, but it is not certain that the field during a transition is that of a dipole. Detailed analyses of field behavior during a polarity transition usually show a notable decrease in field intensity (Fig. 12.27); this is observed in volcanic and sedimentary records of reversals. Possibly the dipole component disappears, granting more importance to higher-order quadrupole or octupole field configurations. On the other hand, it is possible that the transitional field remains dominantly that of a dipole. If this can be assumed, the position of the virtual geomagnetic pole (VGP) of the transitional dipole field can be calculated. During a polarity transition the VGP position changes progressively. It appears to move systematically relative to the Earth's rotation axis, defining a path from one polar region to the opposite one.

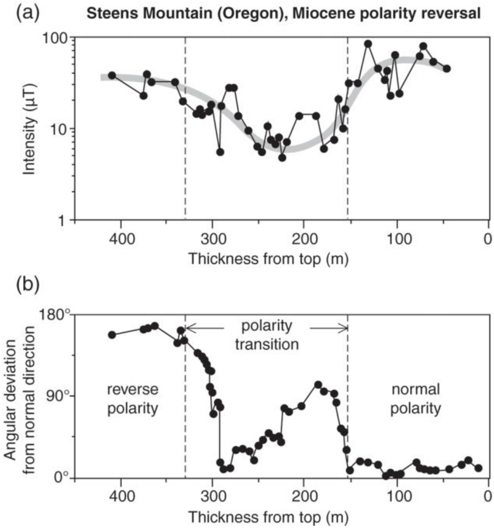


Fig. 12.27 The record of a reversed-to-normal Miocene polarity transition at Steens Mountain, Oregon. (a) The paleointensity record during the transition and (b) the directional record, shown as the angular deviation from the normal paleomagnetic direction outside the transition.

Source: M. Prévot, E.A. Mankinen, C.S. Grommé and R.S. Coe, How the geomagnetic field vector reverses polarity, *Nature* 316, 230–234, 1985.

12.3.1.2 Geomagnetic Polarity Intervals

Long intervals of constant normal or reversed polarity, originally called polarity *epochs*, are referred to as *polarity chrons* (Fig. 12.28); they last typically from 50 kyr to 5 Myr. The polarity chrons are interrupted at irregular intervals by shorter *polarity subchrons* (originally called *events*) lasting for 20–50 kyr. At times the polarity record shows large departures of the magnetic pole from normal or reversed polarity, but the polarity does not change completely; the pole wanders into equatorial latitudes but returns to its initial location on the rotation axis. The departure is short-lived, lasting less than 10 kyr, and the phenomenon is called a magnetic *excursion*. The nature of these features is controversial. The most common interpretation is that they are attempted reversals that were not completed. An alternative interpretation is that they are complete polarity reversals that lasted only a few thousand years, in which case they may only be fully recorded in high deposition rate sediments.

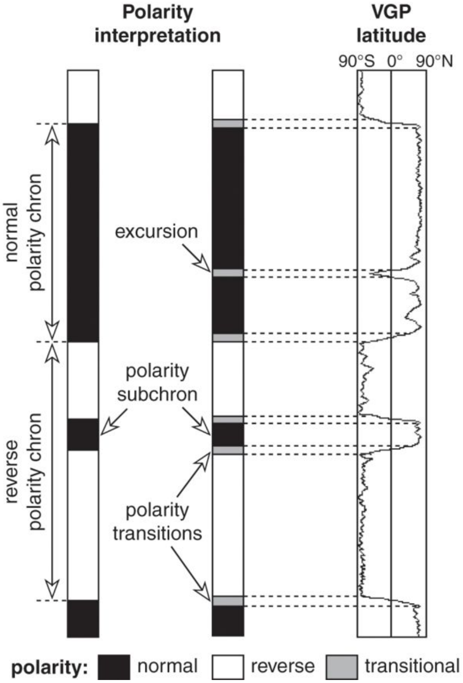

Polarity interpretation
VGP latitude
90°S
0°
90°N
normal polarity chron
reverse polarity chron
excursion
polarity subchron
polarity transitions
polarity:
normal
reverse
transitional

Fig. 12.28 Definition of polarity chrons, subchrons, and transitions.

Source: W.B. Harland, R.L. Armstrong, A.V. Cox, L.E. Craig, A.G. Smith and D.G. Smith, *A Geologic Time Scale 1989*, 263 pp., Cambridge, England: Cambridge University Press, 1990.

The irregular pattern of polarity intervals in a sequence provides a kind of geological fingerprint that can be used under favorable circumstances to date and correlate some types of sedimentary rocks. This procedure is called magnetic polarity stratigraphy, or *magnetostratigraphy*.

12.3.2 Magnetostratigraphy in Lavas and Sediments

In the 1950s the methods of dating rocks took a giant step forward with an improvement in techniques for dating rocks radiometrically. The potassium–argon method (Section 8.4.4) was used to measure accurate ages for Pliocene and Pleistocene lava samples from flows that were also sampled for paleomagnetic purposes. The polarity of the TRM of the lava was found to correlate with its age. There were distinct intervals of time in which the field polarity was the same as at present, and these were separated by intervals of exactly opposite polarity. At first, data were sparse, and in the earliest interpretations it was thought that the field changed polarity quite regularly, roughly once every million years (Fig. 12.29). Gradually, however, a more complex history evolved. Long *epochs* (now referred to as *chrons*) of a given polarity were found to contain much shorter *events* (or *subchrons*) of opposite polarity. The polarity chrons were named after important investigators of paleomagnetic polarity (Brunhes, Matuyama) and geomagnetism (Gauss, Gilbert) while the polarity events were named after the geographical location where they were first discovered (Jaramillo Creek in New Mexico, Olduvai Gorge in Africa, etc.). Countless studies of magnetic polarity in radiometrically dated lavas have established the history of geomagnetic polarity during the last 5 Myr. The dated polarity sequence is called a *geomagnetic polarity timescale*.

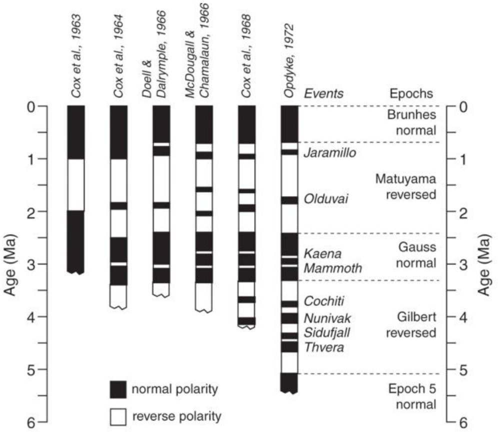


Fig. 12.29 Progressive evolution and refinement of the magnetic polarity timescale.

There is a practical limit to the application of this technique. As can be seen by quick inspection, some of the polarity events last less than 50 kyr. If a reasonable precision of 1–2% is assumed for potassium–argon dating, the error in determining the age of a lava sample that is about 5 Myr old amounts to 50–100 kyr. This is longer than the duration of many short events. The dating error makes it impossible to associate the lava sample unambiguously with the correct polarity event. Extension of the magnetic polarity timescale beyond 5 Myr requires resorting to other methods.

In the mid-1960s the polarity record from lavas was augmented by a large amount of high-quality data acquired from young deep-sea sediments. The deep-ocean basins provide a tranquil depositional environment, where sediments are deposited at rather uniform rates. Marine geologists routinely take cores of sediment for sedimentological, geophysical, and paleontological studies using a special coring device (an example is shown in Fig. 9.14). The magnetostratigraphy of a core is studied by measuring the direction of magnetization in small oriented samples from different depths in the core. Although the axes of deep-sea cores are vertical, they are not oriented in azimuth, so the declinations can only be determined relative to an arbitrary reference value. Thus polarity determinations are often based only on the inclination records. The boundaries between normal and reverse magnetozones are interpolated at the depths where the inclination is zero (Fig. 12.28). In equatorial cores the inclinations are nearly zero, but relative changes in declination can show the polarity record.

The polarity records of numerous deep-sea cores correlate well with the polarities found in contemporaneous young lavas. The deep-sea magnetic polarity records are independent of sedimentary lithology; the same reversal occurs at different depths from core to core because of different sedimentation rates (Fig. 12.30). The magnetic reversals can be used to assign absolute ages to a sediment. The reversal depths in the sediment are correlated directly with the radiometrically dated lava record, from which the incremental sedimentation rates in the sediment core are then calculated. A dated magnetic polarity stratigraphy gives the absolute ages of the first and last appearances of fossils, and so allows absolute dating of paleontological fossil zones.

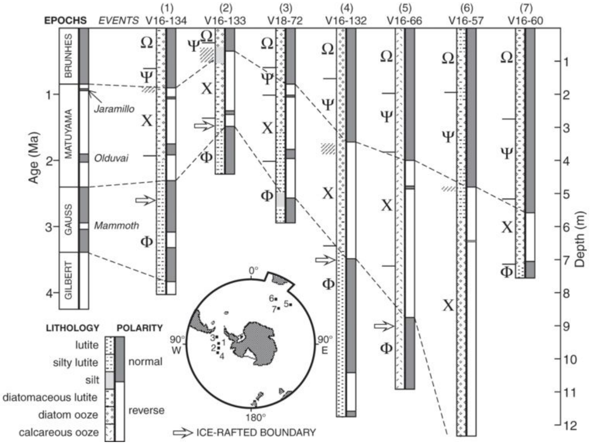

EPOCHS
EVENTS
(1) V16-134
(2) V16-133
(3) V18-72
(4) V16-132
(5) V16-66
(6) V16-57
(7) V16-60
BRUNHES
MATUYAMA
GAUSS
GILBERT
Jaramillo
Olduvai
Mammoth
Age (Ma)
1
2
3
4
Ω
Ψ
X
Φ
Depth (m)
1
2
3
4
5
6
7
8
9
10
11
12
LITHOLOGY
lutite
silty lutite
silt
diatomaceous lutite
diatom ooze
calcareous ooze
POLARITY
normal
reverse
ICE-RAFTED BOUNDARY
0°
90° W
90° E
180°
1
2
3
4
5
6
7

Fig. 12.30 Magnetic reversals in Antarctic deep-sea sediment cores correlate with the radiometric polarity timescale. This allows fossil zones (Greek letters) to be dated. Tie-lines between reversals illustrate the effects of different sedimentation rates.

Source: N.D. Opdyke, B. Glass, J.D. Hays and J. Foster, Paleomagnetic study of Antarctic deep-sea cores, *Science* 154, 349–357, 1968.

A more accurate way of dating sedimentary sequences is the modern technique of *cyclostratigraphy*. This consists of the identification and counting of Milankovitch cycles (Section 3.3.5.5) in a sedimentary section. It has become a valuable technique for precisely dating sediments, provided these contain cyclical changes of sedimentary properties. These properties may be optical (variations in color), stratigraphic (bedding thicknesses), or geochemical (isotopic ratios). In sediments that have a magnetic polarity record, cyclostratigraphy yields more exact measures of the ages and durations of polarity zones. In turn, correlation with the marine polarity record gives more exact dating of marine magnetic anomalies and the ocean floor (Section 12.3.4.1).

Cyclostratigraphy was initially combined with magnetic stratigraphy to date magnetic reversals in young sediments from the Mediterranean that contained sapropels, producing an astronomically tuned timescale for the Late Neogene. The combination was very successful and has been applied subsequently to older sediments, in particular from cores drilled in the deep-ocean floor. Gradually the method has been extended into older sedimentary sequences in the Mesozoic, where the cyclostratigraphy determines the duration of a time interval. A polarity timescale results by combining the cyclostratigraphy with radiometric dates. A case study in the Triassic is described in Section 12.3.5.

The magnetostratigraphic data from deep-sea sediment cores eliminated the lingering doubt that reversals observed in lavas may be due to a self-reversal mechanism. Lavas and sediments acquire their magnetizations by quite different mechanisms; the TRM of a lava is acquired rapidly during cooling from high temperature, whereas the depositional or post-depositional remanent magnetization in a sediment is acquired slowly at constant ambient temperature. Although a self-reversal mechanism might be invoked to cast doubt on polarity changes in lavas, the argument is invalid for the remanence of deep-sea sediments. The common polarity sequence in lavas and sediments can only be explained as a record of the alternations of polarity of the Earth's magnetic field. Moreover, the same pattern of reversals is found regardless of geographical location, emphasizing that reversals are a global phenomenon.

12.3.3 Marine Magnetic Anomalies and Geomagnetic Polarity History

The striped magnetic anomaly patterns formed at oceanic ridges contribute to the compelling geophysical evidence in favor of the theory of global plate tectonics (Section 2.5). Marine magnetic surveys and independent investigations of the rock magnetic properties of marine rocks and sediments have identified the source of the magnetic anomalies to be the basaltic Layer 2 A of the oceanic crust.

Seismic evidence indicates that the oceanic crust has a vertically stratified structure (Fig. 7.38). The uppermost part, seismic Layer 1, consists of a layer of slowly accumulating marine sediments; the thickness of the sediments increases progressively away from the ridge crest. The sediments are so weakly magnetic that they are essentially transparent to the Earth's magnetic field. The seismic Layer 2 A consists of a 500 m thick layer of oceanic basalts that are extruded as submarine lava flows or intruded as dikes. These basalts are strongly magnetic and are chiefly responsible for the strong magnetic anomalies observed at the ocean surface. The metamorphosed basalts of the underlying Layer 2B are too weakly magnetic to have much signature. The rocks of the deeper gabbroic Layer 3 may be sufficiently magnetic to add to the skewness of the magnetic anomalies.

12.3.3.1 Marine Magnetic Anomalies

The origin of marine magnetic anomalies is explained by the Vine–Matthews–Morley hypothesis (Section 2.5.1). Oceanic basalts are found to have strong and stable remanent magnetizations. Their Königsberger ratios are much larger than unity, so the remanent magnetizations are more important than the magnetization induced by the present-day geomagnetic field. The conventional method of interpreting surveyed magnetic anomalies assumed that the anomaly was due to the *susceptibility contrast* between adjacent crustal blocks. According to the Vine–Matthews–Morley hypothesis the oceanic magnetic anomalies arise from the contrast in *remanent magnetizations* between adjacent, oppositely magnetized crustal blocks. The remanent magnetization is acquired thermally by the basalts in oceanic crustal Layer 2A.

The main magnetic mineral in oceanic basalts is titanomagnetite (Section 12.1.3.1). A basaltic lava is initially at a temperature well above 1000°C. Its titanomagnetite grains frequently have skeletal structures, indicating that they cooled and solidified so rapidly that there was not enough time for the formation of normal crystals. Eventually the temperature of the lava sinks below the Curie point of the titanomagnetite (around 200–300°C) and the lava acquires a TRM in the direction of the Earth's magnetic field at that time. Basalts formed contemporaneously along an active spreading ridge acquire the same polarity of magnetization. Long, thin strips of similarly magnetized crust form on opposite sides of the spreading center. These elongated "crustal blocks" may be hundreds of kilometers in length parallel to the ridge axis and several tens of kilometers wide normal to the ridge, while Layer 2 A – the strongly magnetic upper part – is only 0.5 km thick.

Sea-floor spreading persists for millions of years at an oceanic ridge. During this time the magnetic field changes polarity many times. The alternating field polarity leaves some blocks of oceanic crust normally magnetized while their neighbors are reversely magnetized. When the total intensity of the field is measured from a survey ship or aircraft, an alternating sequence of positive and negative anomalies is observed (Fig. 2.10), which can be interpreted in terms of the crustal magnetization (Section 11.4.6). The anomalies can be correlated almost linearly between parallel profiles across a ridge system; consequently, the stripe-like anomalies are often referred to as *magnetic lineations*.

12.3.3.2 Uniformity of Sea-Floor Spreading

Each anomaly in a set of magnetic lineations derives from a crustal block (or stripe) that formed at a ridge and was subsequently transported away from the spreading center. A magnetized crustal block forms during a period of sea-floor spreading when the geomagnetic polarity was constantly normal or reversed, and therefore represents a polarity chron or subchron. The width of a particular block depends on the duration of the chron and the spreading rate at the ocean ridge.

The rate at which the sea-floor is spreading can be determined directly close to a ridge (Section 2.5.2). The edges of the magnetized crustal stripes correspond to the occurrences of polarity reversals, which can be correlated directly with the radiometrically dated sequence for the last 3–4 Myr determined in lavas on the continents or islands. A plot of the distance of a given polarity reversal from the spreading axis against the age of the reversal is nearly linear near the ridge; the slope of a best-fitting straight line gives the average *half-rate of spreading* at the ridge (Fig. 2.11). This is half the full rate of plate separation, assuming that sea-floor spreading has been symmetric on each side of the ridge, which is often the case.

Accumulated evidence from marine magnetic profiles allows us to assess the constancy of sea-floor spreading at different spreading ridge systems. It is thought to have been uniform for the longest time in the South Atlantic. A plot of the distance to a given anomaly in the South Atlantic against the distance to the same anomaly in the Indian, North Pacific, and South Pacific Oceans contains several long linear segments, representing constant rates of sea-floor spreading in both oceans defining the line (Fig. 12.31). A change in gradient indicates a change in spreading rate in one ocean relative to the other. The plot does not exclude a synchronous worldwide change of spreading rate in all oceans, but this would be a rather unlikely occurrence. Clearly the rate of sea-floor spreading changes from time to time, but for long intervals it is a remarkably constant process.

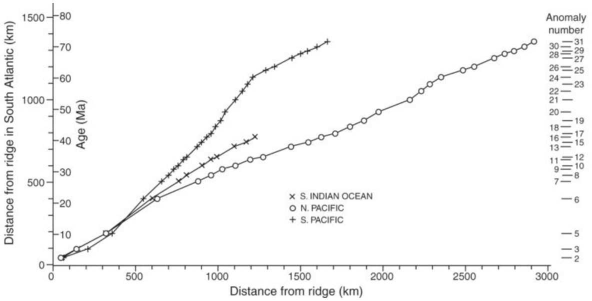


Fig. 12.31 Distances of anomalies from the ridge axis in the South Atlantic plotted against distances to the same anomalies from spreading centers in the Indian and North and South Pacific Oceans.

Source: J.R. Heirtzler, G.O. Dickson, E.M. Herron, W.C. Pitman III and X. Le Pichon, Marine magnetic anomalies, geomagnetic field reversals, and motions of the ocean floor and continents, *J. Geophys. Res.* 73, 2119–2136, 1968.

12.3.3.3 The Marine Record of Geomagnetic Polarity History

Investigations of magnetic anomalies in all major oceanic areas have given a clear, consistent record of the history of geomagnetic polarity during the past 155–160 Myr. It consists of two sequences of polarity reversals represented by magnetic lineations and a long interval of constant normal polarity (Fig. 12.32). The sequences of chrons derived from magnetic lineations have been confirmed by magnetostratigraphic research.

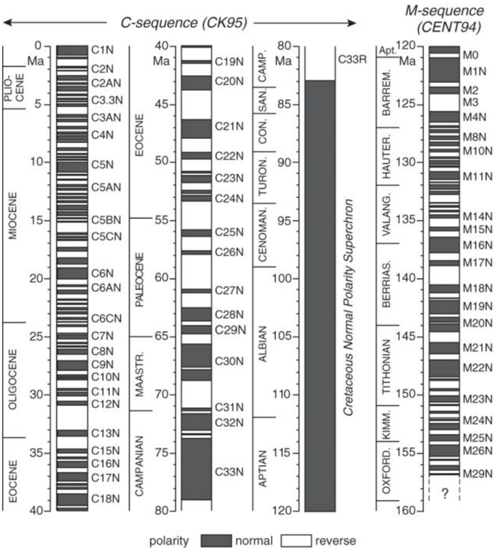


Fig. 12.32 The geomagnetic polarity timescale since the Late Jurassic, derived from the interpretation of marine magnetic anomalies and calibrated by coordinated magnetostratigraphy and biostratigraphy. The polarity record of the C-sequence anomalies for the past 85 Ma (CK95) is from Cande and Kent (1992, 1995); the M-sequence record from 120 Ma to 157 Ma (CENT94) is that of Channell et al. (1995). The record of reversals prior to about 155 Ma is uncertain.

The most prominent positive magnetic anomalies are numbered in turn from the youngest (anomaly 1 at an active ridge axis) to the oldest (anomalies 33–34 in the Late Cretaceous). The associated polarity chrons are identified by the same number and a letter to indicate the polarity. The polarity chrons in the latest sequence fall largely in the Cenozoic (Fig. 8.2) and are identified by a leading letter C, which may be taken to stand for either “chron” or “Cenozoic.” The present Brunhes normal polarity interval corresponds to chron C1N; the reversed interval older than it is labeled chron C1R. Anomaly 2 corresponds to the normal Olduvai event, which interrupts the reversed Matuyama interval and is identified as polarity chron C2N; the reversed interval older than C2N is called polarity chron C2R, etc.

The current episode of magnetic reversals began in the Late

Cretaceous. The oldest normal polarity chron in the sequence is C33N; it is preceded by reversed polarity chron C33R, which ended a long interval (lasting about 35 Ma) in which no polarity reversals took place. This interval in which the geomagnetic field had a constant normal polarity is variously called the Cretaceous Quiet Interval, the Cretaceous Normal Polarity Superchron, or chron C34N.

A phase of alternating polarity giving lineated magnetic anomalies precedes the Cretaceous Quiet Interval. It began late in the Middle Jurassic and continued until the middle of the Early Cretaceous. These Late Mesozoic oceanic anomalies are referred to as the M-sequence. To distinguish them from the later sequence the numbered chrons are identified by a leading letter M. The youngest anomalies M0–M8 are numbered sequentially regardless of the magnetization polarity; older than M9 normal and reverse polarity chrons are paired, as for the Cenozoic. The oldest securely identified chron in the sequence is M29N. Older oceanic crust was once regarded as a Jurassic Quiet Zone that formed when no reversals were taking place. However, lineated anomalies with low amplitudes have been identified in these oceanic regions and interpreted as polarity chrons. It has been difficult to locate them in terrestrial sections and thus it has not yet been established that they represent polarity reversals rather than geomagnetic intensity fluctuations.

The oldest regions of the modern oceans correspond to oceanic crust formed approximately 175 Myr ago, when Pangea broke up and the current episode of sea-floor spreading was initiated. The marine magnetic anomalies over these areas have subdued amplitudes and they do not form lineations. Either no reversals happened during the period of initial spreading or the oceanic crust has not been able to retain the record. The character of the geomagnetic field during this part of the Early Jurassic has not yet been definitively established.

12.3.4 Geomagnetic Polarity Timescales

The interpretation of marine magnetic anomalies provides the most continuous and reliable record of geomagnetic polarity since the Middle Jurassic. The oceanic record is much longer than the radiometrically dated magnetic polarity timescale, which covers only about the past 5 Myr. Hence, it is not possible to date most of the marine magnetic anomalies by direct correlation with a radiometrically dated polarity sequence. Knowledge of the spreading rate at a ridge system provides an alternative way of determining the age of the oceanic crust. Assuming that the rate of sea-floor spreading at the spreading center is constant, the age of a given anomaly can be computed by dividing its distance from the spreading center by the spreading rate. However, this is an unsatisfactory method because the extrapolation is many times longer than the baseline and does not allow for changes in spreading rate like those documented in Fig. 12.31. An alternative method of dating the oceanic polarity record is by establishing the same polarity sequence in coeval sedimentary rocks that are dated paleontologically. This has been achieved in several magnetostratigraphic investigations in pelagic carbonate rocks. The known absolute ages of major stage boundaries are used as tie-levels, against which the ages of magnetic polarity chrons that are too old to be dated directly are calculated by correlation and interpolation.

12.3.4.1 Magnetostratigraphic Calibration of Marine Polarity Sequences

The marine magnetic anomalies near to a spreading ridge (see Fig. 2.10) were originally dated by correlating their polarity record with the radiometrically dated sequence in continental lavas and paleontological zones in deep-sea sediments. This was restricted to the most recent 5 Myr by the errors in radiometric dating. Independent confirmation and dating of older marine magnetic anomalies has been carried out by combined magnetostratigraphical and paleontological studies in suitable rock formations.

The dating of marine magnetic anomalies by this method is illustrated by a case study in the Umbrian Apennines of Central Italy. The Scaglia Rossa pelagic limestone was deposited almost continuously from the Late Cretaceous to the Eocene. Rock magnetic analysis showed that the limestone contained an easily defined characteristic remanent magnetization. Closely spaced paleomagnetic samples were measured from a 200 m thick section through the limestone exposed in the Bottaccione gorge near Gubbio. The declinations and inclinations of the stable magnetization were determined and, after simple tectonic corrections, were used to calculate the latitude of the virtual geomagnetic pole (VGP) during deposition of the limestone. In times of normal polarity the VGP latitude is near 90°N; during reversed polarity it is 90°S. The fluctuations of VGP latitude clearly define magnetozones of normal and reversed polarity (Fig. 12.33).

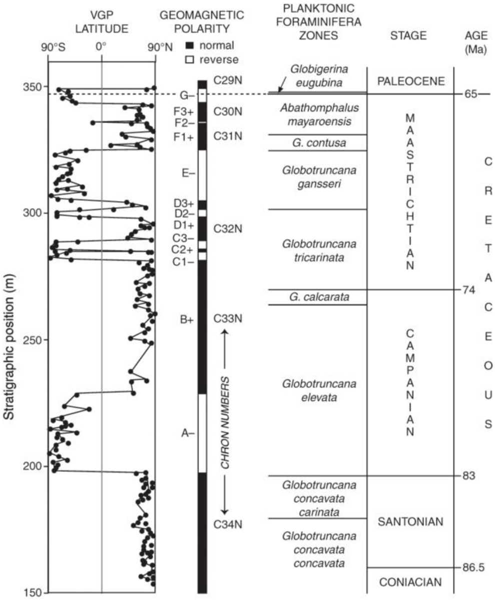


Fig. 12.33 Magnetostratigraphy and biostratigraphy in the Bottaccione section at Gubbio, Italy (after Lowrie and Alvarez,

1977).

The polarity record from the 200 m thick Gubbio limestone section correlates almost perfectly with the polarity sequence derived from marine magnetic anomalies 29–34 measured in different oceans on profiles that are hundreds of kilometers long (Fig. 12.34). The shapes of anomalies 29–34 differ on these profiles because of the different directions of the surveys and the orientations of the spreading axes in the Indian, North Pacific, and South Atlantic Oceans. However, interpretation of the anomalies gives nearly identical crustal magnetization patterns, which agree with the Gubbio polarity record. In addition to confirming the oceanic polarity sequence, the Gubbio results made it possible to date the anomalies by comparison with the geological timescale.

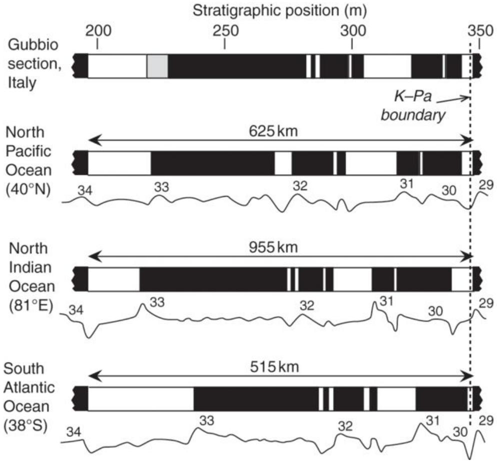

Stratigraphic position (m)
200
250
300
350
Gubbio section, Italy
K–Pa boundary
North Pacific Ocean (40°N)
625 km
34
33
32
31
30
29
North Indian Ocean (81°E)
955 km
34
33
32
31
30
29
South Atlantic Ocean (38°S)
515 km
34
33
32
31
30
29

Fig. 12.34 Crustal magnetization patterns for anomalies 29–34 interpreted from magnetic profiles in the Indian, North Pacific, and South Atlantic Oceans, and comparison with the magnetic polarity stratigraphy in the Gubbio Bottaccione section (after Lowrie and Alvarez, 1977). K–Pa indicates the position of the Cretaceous–Paleogene boundary, which falls within reversed polarity chron C29 R.

Paleontological studies in the Gubbio section gave the locations of important planktonic foraminifera zones and enabled the positions of the major stage boundaries to be located relative to the polarity sequence. The absolute ages of some important stage boundaries are known from independent radiometric and stratigraphic work. This enabled calculation of the absolute ages of the polarity reversals in the Gubbio section, and, by correlation, the ages of the corresponding parts of the ocean floor. For example, the Santonian–Campanian boundary (with an age of about 83 Ma) lies close to the old edge of the reversed polarity chron C33R; the geologically important Cretaceous–Tertiary boundary (age 66 Ma) falls within reversed polarity chron C29R.

In this way the geomagnetic polarity sequence in the Late Cretaceous and Paleogene was confirmed in magnetostratigraphic sections, and the locations of many major and minor stage boundaries were correlated to the polarity sequence paleontologically (Fig. 12.33). Reliable absolute ages for some of the stage boundaries allowed them to be used as calibration levels. The ages of magnetic reversals that fall between the dated tie-points are computed by interpolation, giving a numerical geomagnetic polarity timescale. The method was extended to older overlapping magnetostratigraphic sections of Early Cretaceous and Late Jurassic age, which thus permitted independent confirmation and dating of the polarity record of M-sequence oceanic magnetic anomalies (Fig. 12.32).

12.3.4.2 Frequency of Polarity Reversals

Visual examination of the magnetic polarity record in Fig. 12.32 shows that the rate of polarity reversals has been quite variable in the last 160 Myr. A simple way to portray the reversal rate is to count the numbers of reversals in successive bins of equal duration (e.g., 4 Myr) and compute the average number per million years. Although the reversal rate is uneven, the average frequency during the M-sequence was about 3–4 reversals/Myr. From about 130 Ma the reversal rate slowed down until the beginning of the Cretaceous Normal Polarity Superchron (CNPS) at about 121 Ma. There were no reversals for the following 38 Myr. The CNPS ended at about 83 Ma, after which the reversal rate resumed slowly and then gradually speeded up. It reached a peak of about 5 reversals/Myr in the Late Miocene, about 10 Myr ago, and has since decreased slightly to about 4 reversals/Myr. The current interval of normal polarity (known as the Brunhes or chron C1N) has lasted for 0.78 Myr, which is much longer than the mean length of polarity chrons in the Late Cenozoic.

12.3.4.3 Reconstruction of Plate Tectonic Motions

Once the ages of magnetic anomalies are known, a map showing the positions of dated anomalies is equivalent to a chronological map of the ocean basins (Fig. 12.35). The different rates of sea-floor spreading are evident from the separations of the isochrons. The oldest domains of the oceans have ages about 175 Ma and are found in the Atlantic Ocean adjacent to the coastlines of North America and Africa, as well as in the North Pacific Ocean. They are very much younger than the oldest rocks from the continents, which can have ages up to 3.4 Ga. The oceanic crust has been entirely produced since the onset of sea-floor spreading and the anomaly patterns reflect plate motions.

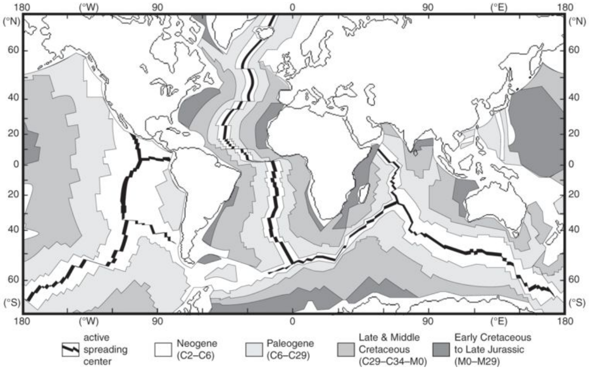

180
(°W)
90
0
90
(°E)
180
(°N)
60
40
20
0
20
40
60
(°S)
active spreading center
Neogene (C2–C6)
Paleogene (C6–C29)
Late & Middle Cretaceous (C29–C34–M0)
Early Cretaceous to Late Jurassic (M0–M29)

Fig. 12.35 Simplified map of the age of oceanic crust, as interpreted from marine magnetic anomalies.

Source: C.R. Scotese, L.M. Gahagan and R.L. Larson, Plate tectonic reconstructions of the Cretaceous and Cenozoic ocean basins, *Tectonophysics* 155, 27–48, 1988.

The past motions of the major plates can be obtained in some detail from the anomaly ages given by a geomagnetic polarity timescale. The relative positions of the continents at any time since the late Mid-Jurassic can be reconstructed by matching coeval marine magnetic anomalies formed at the same spreading center. The procedure is similar to the reconstruction of supercontinents by matching coastlines, 500 fathom depth contours, or APW paths.

The method is illustrated by the sea-floor spreading between North America and Africa in the Central Atlantic (Fig. 12.36). Anomaly 33 forms a long stripe on each side of the Mid-Atlantic Ridge and subparallel to it. The anomaly is due to the magnetization contrast between chron C33N and the reverse polarity chron C33R, which marks the end of the Cretaceous Quiet Interval. Its age is about 81 Ma. The anomaly on the west side of the ridge was formed at the same time as the anomaly on the east side, when the newly formed crust was magnetized at the ridge axis. If the African and North American plates are moved closer together until the east and west anomalies overlap, or until they match along their lengths with minimum misfit, the continents will be brought into the same positions relative to each other that they occupied around 81 Myr ago. By repeating this process of matching dated anomalies it is possible to reconstruct the successive relative positions of the African and North American plates as they separated.

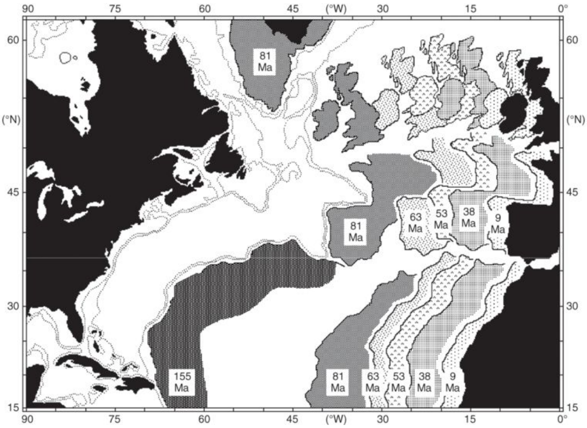


Fig. 12.36 Reconstruction of the history of opening of the North and Central Atlantic oceans. The figure shows the relative positions of Europe and Africa with respect to North America at specific times before the present.

Source: W.C. Pitman III and M. Talwani, Sea-floor spreading in the North Atlantic, *Geol. Soc. Amer. Bull.* 83, 619–646, 1972.

Marine magnetic anomalies in the North Atlantic can be used likewise to describe the relative plate motions between Europe and North America. A picture evolves of the separate histories of separation of Africa and Europe from North America. The differences between the European–North American and African–North American plate motions permit the history of relative motion between Africa and Europe to be inferred. In the Late Cretaceous and Early Tertiary, Africa moved eastwards in a giant shear motion relative to Europe, but since the Middle Tertiary the motion of Africa has been one of convergence and collision with the European plate. This is compatible with the formation of the Alpine fold-belt and the present-day seismicity in the Alpine region.

12.3.5 Early Mesozoic and Paleozoic Reversal History

The marine magnetic anomaly record has excellent resolution; it has made it possible to construct a well-dated history of geomagnetic polarity for about the last 155 Myr. The determination of a geomagnetic polarity timescale for eras older than the Middle Jurassic is more complicated, because no comparable oceanic record exists. The polarity record must therefore be pieced together from continental magnetostratigraphic sections.

A dated polarity record for the Late Triassic and Early Jurassic was obtained in this way in a multidisciplinary study of continental strata from the Newark Basin in the eastern USA by D. V. Kent and colleagues. Stable paleomagnetic directions were determined in more than 2000 paleomagnetic samples taken from more than 6000 meters of core drilled through the sediments. They formed a clear and unambiguous sequence of alternating geomagnetic polarity. Climatically induced variations in lake levels modulated the deposition of the sediments and produced cyclical changes in lamination and color, the most significant of which is called the McLaughlin cycle. This corresponds to the 405 kyr Milankovitch variation in eccentricity of the Earth’s orbit. Cyclostratigraphy (Section 12.3.2) based on the McLaughlin cycle was used to date the lacustrine sediments. Anchored by U–Pb zircon dates at the young end of the sequence, the project produced a geomagnetic polarity timescale for the Late Triassic and earliest Jurassic, spanning estimated ages from 199 Ma to 233 Ma.

No detailed polarity sequence has yet been established earlier than the Late Triassic, although numerous paleomagnetic studies have demonstrated normal and reverse polarity magnetizations in older rocks. As yet, it has only been possible to analyze Paleozoic polarity history in terms of the bias toward normal or reverse polarity. When as many reversed as normal magnetizations are found it is assumed that the field polarity has been reversing; the interval is called a mixed polarity superchron. The Late Cretaceous and Cenozoic and the Triassic illustrate mixed polarity superchrons. Sometimes, for unknown reasons, the polarity was constant for long intervals. This was the case during the Cretaceous Quiet Interval, which in terms of polarity bias is called the Cretaceous Normal Polarity Superchron. The Early Permian and Late Carboniferous were together dominated by reverse polarity. A long interval of constant reverse polarity – the Kiaman interval – corresponds to a superchron referred to as the Permo-Carboniferous Reversed Polarity Superchron. Earlier in the Paleozoic, magnetic reversals were common, but no unique record of the polarity sequence is available.

The derivation of a detailed history of geomagnetic polarity for Early Mesozoic and Paleozoic time is an ongoing task for paleomagnetists and biostratigraphers. The absence of a marine magnetic record before the Middle Jurassic, when the present phase of sea-floor spreading began, requires that earlier polarity sequences must be confirmed by repetition in coeval magnetostratigraphic sections in order to be accepted as globally significant and representative of dipole field behavior.

12.4 *Suggestions for Further Reading*

Introductory Level

Butler, R. F. 1992. *Paleomagnetism: Magnetic Domains to Geologic Terranes*. Boston, MA: Blackwell Scientific.

Intermediate and Advanced Level

Dunlop, D. J. and Özdemir, Ö. 1997. *Rock Magnetism: Fundamentals and Frontiers*. Cambridge: Cambridge University Press.

Evans, M. E. and Heller, F. 2003. *Environmental Magnetism: Principles and Applications of Enviromagnetics*. New York: Academic Press.

Opdyke, N. D. and Channell, J. E. T. 1996. *Magnetic Stratigraphy*. San Diego, CA: Academic Press.

Tauxe, L. 2002. *Paleomagnetic Principles and Practice*. Dordrecht: Kluwer Academic Publishers.

Van der Voo, R. 1993. *Paleomagnetism of the Atlantic, Tethys and Iapetus Oceans*. Cambridge: Cambridge University Press

12.5 Review Questions

1. What are diamagnetism, paramagnetism, ferromagnetism, ferrimagnetism, and antiferromagnetism?

2. Which type of magnetism is exhibited by: (a) quartz, (b) calcite, (c) clay minerals, (d) magnetite, and (e) hematite?

3. What is remanent magnetization? What types of remanent magnetization are possible in: (a) sedimentary, (b) metamorphic, and (c) igneous rocks? Explain how they originate.

4. Explain what is wrong with the following explanation of the origin of thermoremanent magnetization (TRM): "When cooled below the Curie point the magnetite grains align with the magnetic field to give a TRM."

5. What are (a) the declination and (b) the inclination of the magnetic field at a given location? What is the direction of the magnetic field at the magnetic equator?

6. What is implied if the inclination of the remanent magnetization of a rock differs from the inclination of the magnetic field at the same location? What is implied if the declinations differ?

7. How might a rock acquire more than one component of remanent magnetization?

8. How can the original direction of remanent magnetization be identified in a rock that has more than one magnetization component?

9. What is the *geocentric axial dipole* (GAD) hypothesis and why is it important for paleomagnetism?

10. What is an apparent polar wander (APW) path? Why is it an *apparent* path? How are APW paths for different continents interpreted in terms of global tectonics?

11. What is magnetic polarity stratigraphy? How is magnetic polarity stratigraphy calibrated for young rocks and for old rocks?

12. How do oceanic magnetic anomalies originate? Explain the survey procedure that might be used to measure them.

13. What is a geomagnetic polarity timescale? How is a geomagnetic polarity timescale constructed and calibrated?

12.6 Exercises

1. Assuming that the geomagnetic field corresponds to a geocentric axial dipole, calculate the latitude of a site where the field inclination is 45°.

2. With the same assumption, calculate the inclination of the geocentric axial dipole field at latitude 45°N.

3. The north (*N*), east (*E*), and vertically downward (*V*) components of the magnetization of a rock can be calculated from its intensity (*M*), declination (*D*), and inclination (*I*) using the following relationships:

$$\begin{aligned} N &= M \ \cos I \cos D \\ E &= M \ \cos I \sin D \\ V &= M \ \sin I. \end{aligned}$$

In the progressive thermal demagnetization of a sample of Cretaceous limestone for a paleomagnetic study, the following remanent magnetizations were measured at different temperatures *T*:

T (°C)	*M* (10^{-5} A/m)	*D* (°)	*I* (°)
200	5.08	60.8	60.1
300	3.87	62.1	59.8
400	3.02	61.2	62.1
500	2.10	62.2	61.6
550	1.18	63.1	60.9

(a) Calculate the north, east, and vertical components of the magnetizations at each demagnetization stage.

(b) Plot the *N*-components against the *E*-components, fit a straight line, and determine the optimum declination for the stable magnetization direction.

(c) Plot the V-components against either the N- or E-components, fit a straight line, and compute the optimum inclination for the stable magnetization direction.

(d) The straight lines do not pass through the origin of the plot. What does this imply?

4. In the same paleomagnetic study, the following stable directions of remanent magnetization, corrected for the dip of bedding in the limestone formation, were measured in five samples from the same site:

Sample	***D*** **(°)**	***I*** **(°)**
SR-04	329.7	40.6
SR-05	336.6	24.7
SR-07	326.2	46.0
SR-10	321.1	40.9
SR-12	322.7	44.9

(a) Calculate the direction cosines (λ_N, λ_E, λ_V) of the north, east, and vertical (V) components of each stable direction, using the relationships

$$\begin{aligned}\lambda_N &= \cos I \cos D\\ \lambda_E &= \cos I \sin D\\ \lambda_V &= \sin I.\end{aligned}$$

(b) Add up the values for each direction cosine. Let the sums be X, Y, and Z, where

$$X=\sum\lambda_N;\ Y=\sum\lambda_E;\ Z=\sum\lambda_V.$$

Calculate the vector sum of the directions, R, and the declination D_m and inclination I_m of the mean direction using the relationships

$$R = \sqrt{X^2 + Y^2 + Z^2}; \tan D_m = Y/X; \ \sin I_m = \ Z/R.$$

(c) Using the computed value of R, calculate the precision parameter k of the data and the 95% confidence error (α_{95}) for the mean direction.

5. The samples analyzed in the previous exercise were gathered at a site in Italy with latitude 43.4°N, longitude 12.6°E. Calculate the latitude and longitude of the paleomagnetic pole position for the Italian limestone.

6. Assume that the location of the Late Cretaceous paleomagnetic pole for the European plate is at 72°N, 154°E and the corresponding pole for the African plate is at 67°N, 245°E.

(a) Calculate the expected "European" and "African" directions at the Italian site in the previous exercise.

(b) Compare the expected directions with the observed directions and explain how the paleomagnetic results from the Italian limestone should be interpreted.

7. In a paleomagnetic reversal test the normally magnetized samples have a mean direction $D_N = 313°$, $I_N = 38°$, and the circle of 95% confidence has radius $\alpha_{95} = 8°$. The reversely magnetized samples have $D_R = 114°$, $I_R = -33°$, and $\alpha_{95} = 7°$. Calculate the angle between the normal and reverse mean directions. Do these directions pass the reversals test?

12.7 Computational Exercises

Jupyter notebook computational exercises can be found online for this chapter at www.cambridge.org/FoG3ed (Apparent polar wander).

Appendix A

The Three-Dimensional Wave Equation

General Equations of Motion

The development of the equations for P-waves and S-waves in three dimensions begins like the one-dimensional cases. In this case, however, the displacement of a particle during the passage of a seismic wave is a vector $\vec{U}$ with components u, v, and w in the x-, y-, and z-directions, respectively.

Consider the forces acting on the faces of the small box shown in Fig. A1. The faces normal to the x-, y-, and z-directions have areas A_x, A_y, and A_z, respectively. Let the force on the left face of area A_x, acting in the x-direction, be $(F_x)_x$ and that on the right face be $(F_x)_x + \Delta(F_x)_x$. If these forces are equal ($\Delta(F_x)_x = 0$) they will cause no motion in the x-direction. If they are unequal, motion will result. If the box is small enough for the change in force to take place uniformly, we can write

$$\Delta(F_x)_x = \frac{\partial(F_x)_x}{\partial x}\Delta x = \frac{\partial(\sigma_{xx}A_x)}{\partial x}\Delta x = \frac{\partial\sigma_{xx}}{\partial x}(\Delta y\Delta z)\Delta x. \tag{A1}$$

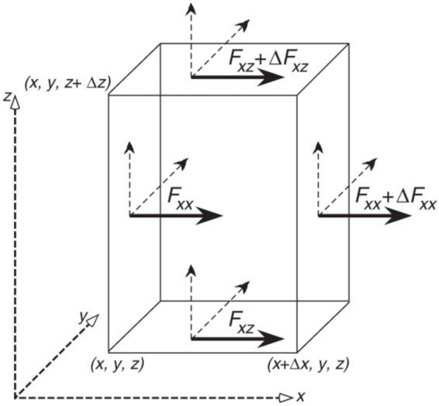


Fig. A1 Normal and shear forces on the faces of a small rectangular box.

The forces acting on the faces normal to the z-direction also have

x-components. Treating them in the same way, the force on the base of the box of area A_z, acting in the x-direction, is $(F_z)_x$ and that on the top face is $(F_z)_x + \Delta(F_z)_x$. The difference is

$$\Delta(F_z)_x = \frac{\partial(F_z)_x}{\partial z}\Delta z = \frac{\partial(\sigma_{zx}A_z)}{\partial z}\Delta z = \frac{\partial\sigma_{zx}}{\partial z}(\Delta x\Delta y)\Delta z. \tag{A2}$$

Analogously (not shown in the figure), the resulting force in the x-direction from the stresses on the faces normal to the y-direction is given by

$$\Delta(F_y)_x = \frac{\partial(F_y)_x}{\partial y}\Delta y = \frac{\partial(\sigma_{yx}A_y)}{\partial y}\Delta y = \frac{\partial\sigma_{yx}}{\partial y}(\Delta z\Delta x)\Delta y. \tag{A3}$$

Combining these results, we get the net force ΔF_x acting on the box in the x-direction:

$$\Delta F_x = \left(\frac{\partial\sigma_{xx}}{\partial x} + \frac{\partial\sigma_{yx}}{\partial y} + \frac{\partial\sigma_{zx}}{\partial z}\right)\Delta V, \tag{A4}$$

where $\Delta V = \Delta x\Delta y\Delta z$ is the volume of the elementary box. If the density of the medium is ρ and the x-component of the displacement is u, we can apply Newton's law of motion to obtain the equation of motion:

$$\left(\frac{\partial\sigma_{xx}}{\partial x} + \frac{\partial\sigma_{yx}}{\partial y} + \frac{\partial\sigma_{zx}}{\partial z}\right)\Delta V = \rho\Delta V\frac{\partial^2 u}{\partial t^2}$$

$$\frac{\partial\sigma_{xx}}{\partial x} + \frac{\partial\sigma_{yx}}{\partial y} + \frac{\partial\sigma_{zx}}{\partial z} = \rho\frac{\partial^2 u}{\partial t^2}. \tag{A5}$$

From the definitions of the Lamé constants (Section 5.1.4.3) and the relationships between stress and strain (Eq. 5.33), we have

$$\sigma_{xx} = \lambda\theta + 2\mu\varepsilon_{xx} = \lambda\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right) + 2\mu\frac{\partial u}{\partial x}$$

$$\sigma_{yx} = 2\mu\varepsilon_{yx} = \mu\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)$$

$$\sigma_{zx} = 2\mu\varepsilon_{zx} = \mu\left(\frac{\partial w}{\partial x} + \frac{\partial u}{\partial z}\right).$$

(A6)

Differentiating each of these equations in turn, as in the equation of motion, gives

$$\frac{\partial \sigma_{xx}}{\partial x} = \lambda\frac{\partial \theta}{\partial x} + 2\mu\frac{\partial^2 u}{\partial x^2}$$

$$\frac{\partial \sigma_{yx}}{\partial y} = \mu\left(\frac{\partial^2 v}{\partial y \partial x} + \frac{\partial^2 u}{\partial y^2}\right)$$

$$\frac{\partial \sigma_{zx}}{\partial z} = \mu\left(\frac{\partial^2 w}{\partial z \partial x} + \frac{\partial^2 u}{\partial z^2}\right).$$

(A7)

After substitution and grouping terms, the equation of motion for the *x*-displacement *u* becomes

$$\lambda\frac{\partial \theta}{\partial x} + 2\mu\frac{\partial^2 u}{\partial x^2} + \mu\left(\frac{\partial^2 v}{\partial y \partial x} + \frac{\partial^2 u}{\partial y^2}\right) + \mu\left(\frac{\partial^2 w}{\partial z \partial x} + \frac{\partial^2 u}{\partial z^2}\right) = \rho\frac{\partial^2 u}{\partial t^2},$$

(A8)

$$\lambda\frac{\partial \theta}{\partial x} + \mu\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 v}{\partial y \partial x} + \frac{\partial^2 w}{\partial z \partial x}\right) + \mu\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right) = \rho\frac{\partial^2 u}{\partial t^2},$$

(A9)

$$\lambda\frac{\partial\theta}{\partial x}+\mu\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}+\frac{\partial w}{\partial z}\right)+\mu\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}+\frac{\partial^2 u}{\partial z^2}\right)=\rho\frac{\partial^2 u}{\partial t^2}.$$

(A10)

It is convenient to introduce the mathematical operator $\nabla^2=\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}$. This allows a more concise presentation of Eq. (A10) for the *x*-displacement, which becomes

$$(\lambda+\mu)\frac{\partial\theta}{\partial x}+\mu\nabla^2 u=\rho\frac{\partial^2 u}{\partial t^2}.$$

(A11)

Similar analysis of the *y*- and *z*-displacements, *v* and *w* respectively, gives the additional equations of motion:

$$(\lambda+\mu)\frac{\partial\theta}{\partial y}+\mu\nabla^2 v=\rho\frac{\partial^2 v}{\partial t^2},$$

(A12)

$$(\lambda+\mu)\frac{\partial\theta}{\partial z}+\mu\nabla^2 w=\rho\frac{\partial^2 w}{\partial t^2}.$$

(A13)

Equations (A11)–(A13) can now be resolved into two wave motions representing longitudinal (P-) waves and transverse (S-) waves.

Longitudinal Waves

Differentiating each of Eqs. (A11), (A12), and (A13) in turn by *x*, *y*, and *z*, respectively, gives

$$(\lambda+\mu)\frac{\partial^2\theta}{\partial x^2}+\mu\nabla^2\left(\frac{\partial u}{\partial x}\right)=\rho\frac{\partial^2}{\partial t^2}\left(\frac{\partial u}{\partial x}\right)$$
$$(\lambda+\mu)\frac{\partial^2\theta}{\partial y^2}+\mu\nabla^2\left(\frac{\partial v}{\partial y}\right)=\rho\frac{\partial^2}{\partial t^2}\left(\frac{\partial v}{\partial y}\right)$$
$$(\lambda+\mu)\frac{\partial^2\theta}{\partial z^2}+\mu\nabla^2\left(\frac{\partial w}{\partial z}\right)=\rho\frac{\partial^2}{\partial t^2}\left(\frac{\partial w}{\partial z}\right).$$

(A14)

Adding these equations together, and using the definitions of ∇^2 given above and $\theta=\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}+\frac{\partial w}{\partial z}$, we obtain

$$(\lambda+\mu)\left(\frac{\partial^2\theta}{\partial x^2}+\frac{\partial^2\theta}{\partial y^2}+\frac{\partial^2\theta}{\partial z^2}\right)+\mu\nabla^2\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}+\frac{\partial w}{\partial z}\right)$$
$$=\rho\frac{\partial^2}{\partial t^2}\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}+\frac{\partial w}{\partial z}\right),$$

(A15)

$$(\lambda+\mu)\nabla^2\theta+\mu\nabla^2\theta=\rho\frac{\partial^2\theta}{\partial t^2},$$

(A16)

$$(\lambda+2\mu)\nabla^2\theta=\rho\frac{\partial^2\theta}{\partial t^2},$$

(A17)

which can be written as

$$\alpha^2 \nabla^2 \theta = \frac{\partial^2 \theta}{\partial t^2}, \tag{A18}$$

where

$$\alpha^2 = \frac{\lambda + 2\mu}{\rho}. \tag{A19}$$

Comparison with the one-dimensional compressional wave equation (Eq. 3.41) shows that Eq. (A18) describes the three-dimensional propagation of a volumetric change (dilatation) θ with velocity α given by Eq. (A19). This wave travels as a succession of dilatations and compressions parallel to the direction of propagation and is referred to as a *longitudinal wave*. Because it is the first to arrive at a seismometer it is more commonly called the *primary wave* or *P-wave*.

Transverse Waves

By differentiating Eq. (A11) with respect to *y*, we get

$$(\lambda+\mu)\frac{\partial^2\theta}{\partial y\partial x}+\mu\nabla^2\left(\frac{\partial u}{\partial y}\right)=\rho\frac{\partial^2}{\partial t^2}\left(\frac{\partial u}{\partial y}\right). \tag{A20}$$

Likewise, differentiating Eq. (A12) with respect to *x* gives

$$(\lambda+\mu)\frac{\partial^2\theta}{\partial x\partial y}+\mu\nabla^2\left(\frac{\partial v}{\partial x}\right)=\rho\frac{\partial^2}{\partial t^2}\left(\frac{\partial v}{\partial x}\right). \tag{A21}$$

Subtracting Eq. (A20) from Eq. (A21), we get

$$\mu\nabla^2\left(\frac{\partial v}{\partial x}-\frac{\partial u}{\partial y}\right)=\rho\frac{\partial^2}{\partial t^2}\left(\frac{\partial v}{\partial x}-\frac{\partial u}{\partial y}\right). \tag{A22}$$

By combining Eq. (A12) with Eq. (A13), and Eq. (A13) with Eq. (A11), in the same way, we obtain the further wave equations

$$\mu\nabla^2\left(\frac{\partial u}{\partial z}-\frac{\partial w}{\partial x}\right)=\rho\frac{\partial^2}{\partial t^2}\left(\frac{\partial u}{\partial z}-\frac{\partial w}{\partial x}\right), \tag{A23}$$

$$\mu\nabla^2\left(\frac{\partial w}{\partial y}-\frac{\partial v}{\partial z}\right)=\rho\frac{\partial^2}{\partial t^2}\left(\frac{\partial w}{\partial y}-\frac{\partial v}{\partial z}\right), \tag{A24}$$

$$\Psi_x = \left(\frac{\partial w}{\partial y} - \frac{\partial v}{\partial z}\right);\ \Psi_y = \left(\frac{\partial u}{\partial z} - \frac{\partial w}{\partial x}\right);\ \Psi_z = \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right).$$
(A25)

The quantities $\frac{\partial v}{\partial z}$, etc., are equivalent to small rotational distortions that combine to give shear strains (see Section 5.1.3.3). Thus, $\left(\frac{\partial v}{\partial z} - \frac{\partial w}{\partial y}\right)$ is a net rotation in the *y–z* plane, i.e., normal to the *x*-direction. The quantities defined in Eq. (A25) are the components of a vector $\Psi = (\Psi_x, \Psi_y, \Psi_z)$. Each component of Ψ describes a net rotation in the plane normal to the specific coordinate axis.

Eqs. (A22)–(A24) show that the components of Ψ satisfy the equation

$$\mu \nabla^2 \Psi = \rho \frac{\partial^2 \Psi}{\partial t^2},$$
(A26)

which can be written as

$$\beta^2 \nabla^2 \Psi = \frac{\partial^2 \Psi}{\partial t^2},$$
(A27)

where

$$\beta^2 = \frac{\mu}{\rho}.$$
(A28)

Equation (A27) has the form of the wave equation. It describes the propagation of a rotational or shear disturbance Ψ with velocity β given by Eq. (A28). Because the displacements in this wave are in the plane normal to the direction of propagation, it is referred to as a *transverse wave*. Its velocity β is slower than the P-wave velocity α, so it arrives at a detector later than the P-wave from the same event and is called the *secondary wave*, or *S-wave*. As the disturbance consists of shear motions of particles in the wavefront, it is also called a *shear wave*.

Appendix B

Cooling of a Semi-Infinite Half-Space

A semi-infinite half-space extending to infinity in the z-direction is initially at a uniform temperature T_0. Let the temperature at time t and depth z inside the cooling half-space be $T(z,t)$. The temperature of the upper surface at $z = 0$ is changed suddenly and maintained at 0°C. The boundary conditions are therefore $T(0,t) = 0$ and $T(z,0) = T_0$. The temperature in the cooling half-space must satisfy the heat conduction equation, which can be written (Box 9.1)

$$\frac{1}{\kappa}\frac{1}{\theta}\frac{d\theta}{dt} = \frac{1}{Z}\frac{d^2Z}{dz^2}. \tag{B1}$$

Separating the variables and setting the separation constant equal to $-n^2$, we get

$$\frac{1}{\kappa}\frac{1}{\theta}\frac{d\theta}{dt} = -n^2 \quad \text{and} \quad \frac{1}{Z}\frac{d^2Z}{dz^2} = -n^2. \tag{B2}$$

The particular solutions of these equations are

$$\theta = \theta_0 e^{-\kappa n^2 t} \quad \text{and} \quad Z = A_n \cos nz + B_n \sin nz. \tag{B3}$$

The boundary condition on the upper surface at $z = 0$ is $T(0,t) = 0$, which requires $A_n = 0$. The general solution can then be written

$$T(z,t) = \theta_0 \sum_n e^{-\kappa n^2 t} B_n \sin nz.$$
(B4)

If the temperature distribution is continuous, the summation can be replaced by an integral and the discrete values (θ_0, B_n) by a continuous function $B(n)$:

$$T(z,t) = \int_0^\infty e^{-\kappa n^2 t} B(n) \sin(nz) \mathrm{d}n.$$
(B5)

If at $t = 0$ the cooling half-space has an initial temperature distribution $T(z)$,

$$T(z,0) = T(z) = \int_0^\infty B(n) \sin(nz) \mathrm{d}n.$$
(B6)

By using the orthogonality properties of sine and cosine functions, we get

$$B(n) = \frac{2}{\pi} \int_0^\infty T(z) \sin(nz) \mathrm{d}z = \frac{2}{\pi} \int_0^\infty T(\zeta) \sin(n\zeta) \mathrm{d}\zeta,$$
(B7)

where the integration variable is changed to avoid confusion. Substituting in Eq. (B5), we get

$$T(z,t) = \frac{2}{\pi} \int_0^\infty T(\zeta) \left[\int_0^\infty e^{-\kappa n^2 t} \sin(nz) \sin(n\zeta) \mathrm{d}n \right] \mathrm{d}\zeta.$$
(B8)

Invoking the trigonometric relationship $2\sin(nz)\sin(n\zeta) = \cos(n(\zeta - z)) - \cos(n(\zeta + z))$ changes the integral in Eq. (B8) to

$$T(z,t) = \frac{1}{\pi}\int_0^\infty T(\zeta)\left[\int_0^\infty e^{-\kappa n^2 t}\left(\cos\left(n(\zeta - z)\right) - \cos\left(n(\zeta + z)\right)\right)\mathrm{d}n\right]\mathrm{d}\zeta. \tag{B9}$$

The integration can be simplified by writing $\kappa t = \alpha,\ (\zeta - z) = u,$ and $(\zeta + z) = v$:

$$T(z,t) = \frac{1}{\pi}\int_0^\infty T(\zeta)\left[\int_0^\infty e^{-\alpha n^2}\cos(nu)\mathrm{d}n - \int_0^\infty e^{-\alpha n^2}\cos(nv)\mathrm{d}n\right]\mathrm{d}\zeta. \tag{B10}$$

Before proceeding with the solution of Eq. (B10), it is first necessary to evaluate the integral

$$Y = \int_0^\infty e^{-\alpha n^2}\cos(nu)\mathrm{d}n. \tag{B11}$$

Note that

$$\frac{\partial Y}{\partial u} = \int_0^\infty \left(-ne^{-\alpha n^2}\right)\sin(nu)\mathrm{d}n. \tag{B12}$$

Integrating Eq. (B12) by parts gives

$$\frac{\partial Y}{\partial u}=\left[\frac{e^{-\alpha n^2}}{2\alpha}\sin(nu)\right]_0^{\infty}-\frac{u}{2\alpha}\int_0^{\infty}e^{-\alpha n^2}\cos(nu)\mathrm{d}n=-\frac{u}{2\alpha}Y, \tag{B13}$$

$$\frac{1}{Y}\frac{\partial Y}{\partial u}=\frac{\partial}{\partial u}ln(Y)=-\frac{u}{2\alpha}, \tag{B14}$$

$$\ln(Y)=-\frac{u^2}{4\alpha}+c=-\frac{u^2}{4\alpha}+\ln(Y_0), \tag{B15}$$

$$Y=Y_0e^{-\frac{u^2}{4\alpha}}. \tag{B16}$$

The constant Y_0 is the value of the integral Y defined in Eq. (B11) for $u = 0$. This integration (evaluated below) has the value

$$Y_0=\int_0^{\infty}e^{-\alpha n^2}\mathrm{d}n=\frac{1}{2}\sqrt{\frac{\pi}{\alpha}}. \tag{B17}$$

Thus,

$$Y=\int_0^{\infty}e^{-\alpha n^2}\cos(nu)\mathrm{d}n=\frac{1}{2}\sqrt{\frac{\pi}{\alpha}}e^{-\frac{u^2}{4\alpha}}. \tag{B18}$$

Inserting this solution in Eq. (B10) gives

$$T(z,t) = \frac{1}{\pi}\int_0^\infty T(\zeta)\frac{1}{2}\sqrt{\frac{\pi}{\alpha}}\left[e^{-\frac{u^2}{4\alpha}} - e^{-\frac{v^2}{4\alpha}}\right]\,\mathrm{d}\zeta$$

$$= \frac{1}{2\sqrt{\pi\kappa t}}\int_0^\infty T(\zeta)\left[e^{-\frac{(\zeta - z)^2}{4\kappa t}} - e^{-\frac{(\zeta + z)^2}{4\kappa t}}\right]\,\mathrm{d}\zeta.$$

(B19)

If the cooling body has initially a uniform constant temperature T_0, then $T(z) = T_0$ and the integration in Eq. (B19) can be written as

$$T(z,t) = \frac{T_0}{2\sqrt{\pi\kappa t}}\left\{\int_0^\infty e^{-\frac{(z-\zeta)^2}{4\kappa t}}\,\mathrm{d}\zeta - \int_0^\infty e^{-\frac{(z+\zeta)^2}{4\kappa t}}\,\mathrm{d}\zeta\right\}.$$

(B20)

In the first integration we can write $w = \frac{(\zeta - z)}{2\sqrt{\kappa t}}$; then $\mathrm{d}w = \frac{1}{2\sqrt{\kappa t}}\mathrm{d}\zeta$ and the upper and lower limits of the integration change to ∞ and $\frac{-z}{2\sqrt{\kappa t}}$, respectively. Similarly, in the second integration we can write $w = \frac{(\zeta + z)}{2\sqrt{\kappa t}}$, with corresponding changes to the integration limits, which become ∞ and $\frac{z}{2\sqrt{\kappa t}}$, respectively. Equation (B20) can be rewritten as

$$T(z,t) = \frac{T_0}{\sqrt{\pi}}\left\{\int_{\frac{-z}{2\sqrt{\kappa t}}}^{\infty} e^{-w^2}\,\mathrm{d}w - \int_{\frac{z}{2\sqrt{\kappa t}}}^{\infty} e^{-w^2}\,\mathrm{d}w\right\},$$

(B21)

$$T(z,t) = \frac{T_0}{\sqrt{\pi}}\left\{\int_{\frac{-z}{2\sqrt{\kappa t}}}^{\frac{z}{2\sqrt{\kappa t}}} e^{-w^2}\,\mathrm{d}w\right\} = \frac{2T_0}{\sqrt{\pi}}\left\{\int_0^{\frac{z}{2\sqrt{\kappa t}}} e^{-w^2}\,\mathrm{d}w\right\},$$

(B22)

$$T(z,t) = T_0\left\{\frac{2}{\sqrt{\pi}}\int_0^{\frac{z}{2\sqrt{\kappa t}}} e^{-w^2}\,\mathrm{d}w\right\}.$$

(B23)

The expression in brackets is the *error function* (Box 4.3), defined as

$$\operatorname{erf}(\eta) = \frac{2}{\sqrt{\pi}}\int_0^{\eta} e^{-u^2}\,\mathrm{d}u.$$

(B24)

The solution for the temperature distribution in the cooling plate is therefore

$$T(z,t) = T_0\operatorname{erf}\left(\frac{z}{2\sqrt{\kappa t}}\right).$$

(B25)

Evaluation of the Integral Y_0 in Eq. (B17)

$$Y_0 = \int_0^\infty e^{-\alpha x^2}\,\mathrm{d}x = \int_0^\infty e^{-\alpha y^2}\,\mathrm{d}y,$$

(B26)

$$(Y_0)^2 = \left(\int_0^\infty e^{-\alpha x^2}\,\mathrm{d}x\right)\left(\int_0^\infty e^{-\alpha y^2}\,\mathrm{d}y\right) = \int_0^\infty\int_0^\infty e^{-\alpha(x^2+y^2)}\,\mathrm{d}x\,\mathrm{d}y.$$

(B27)

This is an integration over an area enclosed by the positive x- and y-axes ($0 \le x \le \infty$; $0 \le y \le \infty$). Change to polar coordinates $x = r\cos\varphi$ and $y = r\sin\varphi$ and write the element of area as $(\mathrm{d}x\,\mathrm{d}y) = (r\mathrm{d}r\,\mathrm{d}\varphi)$. The limits of the integration change to $0 \le r \le \infty$; $0 \le \varphi \le (\pi/2)$:

$$(Y_0)^2 = \int_0^{\pi/2}\int_0^\infty e^{-\alpha r^2} r\mathrm{d}r\,\mathrm{d}\varphi = \int_0^{\pi/2}\left[\int_0^\infty e^{-\alpha r^2} r\,\mathrm{d}r\right]\mathrm{d}\varphi$$
$$= \int_0^{\pi/2}\left[-\frac{e^{-\alpha r^2}}{2\alpha}\right]_0^\infty \mathrm{d}\varphi,$$

(B28)

$$(Y_0)^2 = \int_0^{\pi/2}\left[-\frac{e^{-\alpha r^2}}{2\alpha}\right]_0^\infty \mathrm{d}\varphi = \frac{1}{2\alpha}\int_0^{\pi/2}\mathrm{d}\varphi = \frac{\pi}{4\alpha}.$$

(B29)

The value of the integral in Eq. (B17) is thus

$$Y_0 = \frac{1}{2}\sqrt{\frac{\pi}{\alpha}}.$$

(B30)

Appendix C
Magnetic Behavior of Rock-Forming Minerals

Diamagnetism

The Lorentz law (Box 11.2) shows that a change in the magnetic field acting on an atom alters the force experienced by an orbiting electron. The plane of the electron orbit is compelled to precess around the field direction; the phenomenon is called *Larmor precession*. It represents an additional component of rotation and angular momentum. The sense of the rotation is opposite to that of the orbital rotation about the nucleus. Hence, the magnetic moment associated with the Larmor precession opposes the applied field. As a result, a weak magnetization proportional to the field strength is induced in the opposite direction to the field. The magnetization vanishes when the applied magnetic field is removed. Diamagnetic susceptibility is reversible, weak, and negative (Fig. C1a), and it is independent of temperature. Many important rock-forming minerals belong to this class, among them quartz and calcite. They have susceptibilities around -10^{-6} in SI units.

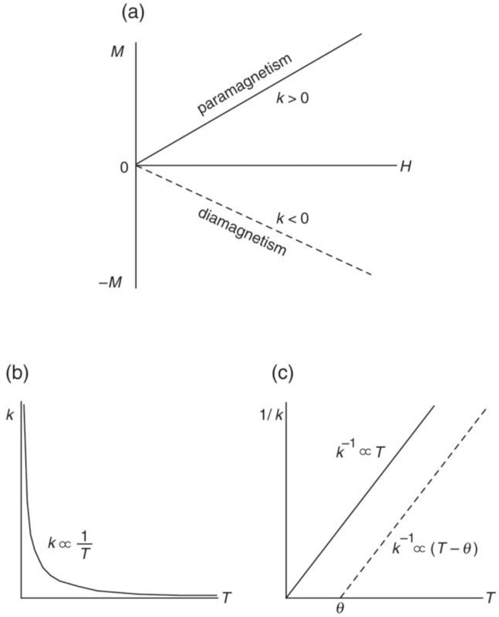


Fig. C1 (a) Variations of magnetization *M* with applied magnetic field H in paramagnetic and diamagnetic materials; (b) the variation of paramagnetic susceptibility with temperature; and (c) the linear plot of the inverse of paramagnetic susceptibility against temperature.

Paramagnetism

When one or more electron spins is unpaired, the net magnetic moment of an atom or ion is no longer zero. The resultant magnetic moment can align with a magnetic field. The alignment is opposed by thermal energy which favors chaotic orientations of the spin magnetic moments. The magnetic energy is small compared to the thermal energy, and in the absence of a magnetic field the magnetic moments are oriented randomly. When a magnetic field is applied, the chaotic alignment of magnetic moments acquires a bias (i.e., a statistical preference) toward the field direction. A magnetization is induced that is proportional to the strength of the applied field and parallel to its direction. The susceptibility is reversible, small, and positive (Fig. C1a). An important paramagnetic characteristic is that the susceptibility k varies inversely with temperature (Fig. C1b) as given by the *Curie law*

$$k = \frac{C}{T}, \qquad (1)$$

where the constant C is characteristic of the material. Thus, a plot of $1/k$ against temperature is a straight line (Fig. C1c).

Interactions between ions in a solid may be quite strong and consequently paramagnetic behavior is only displayed when the thermal energy exceeds a threshold value. The temperature above which a solid is paramagnetic is called the *paramagnetic Curie temperature* or *Weiss constant* of the material, Denoted by θ, it is close to zero kelvin in *paramagnetic* solids. At temperatures $T > \theta$, the paramagnetic susceptibility k is given by the *Curie–Weiss law*

$$k = \frac{C}{T - \theta}.$$

(2)

For a solid the plot of $1/k$ against $(T - \theta)$ is a straight line (Fig. C1c). Many clay minerals and other rock-forming minerals (e.g., chlorite, amphibole, pyroxene, olivine) are paramagnetic at room temperature, with susceptibilities commonly around 10^{-5}–10^{-4} in SI units.

Ferromagnetism

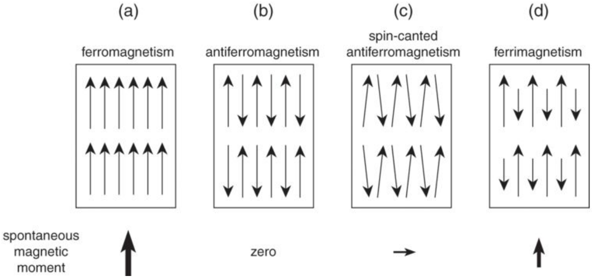

(a)
ferromagnetism
(b)
antiferromagnetism
(c)
spin-canted antiferromagnetism
(d)
ferrimagnetism
spontaneous magnetic moment
zero

Fig. C2 Schematic representations of the alignments of atomic magnetic moments in: (a) ferromagnetism; (b) antiferromagnetism; (c) spin-canted antiferromagnetism; and (d) ferrimagnetism.

In paramagnetic and diamagnetic materials the interactions between individual atomic magnetic moments are small and often negligible. However, in some metals (e.g., iron, nickel, cobalt) the atoms occupy lattice positions that are close enough to allow the exchange of electrons between neighboring atoms. The exchange is a quantum-mechanical effect that involves a large amount of energy, called the *exchange energy* of the metal. The exchange interaction produces a very strong *molecular field* within the metal, which aligns the atomic magnetic moments (Fig. C2a) exactly parallel and produces a *spontaneous magnetization* (M_S). The magnetic moments react in unison to a magnetic field, giving rise to a class of strong magnetic behavior known as *ferromagnetism*. This is characterized by *magnetic hysteresis*, in which the magnetization follows different paths when a magnetizing field is increasing or decreasing (Fig. C3).

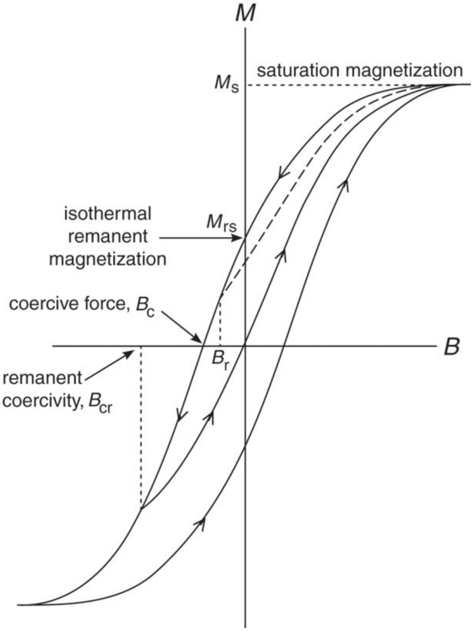


Fig. C3 The hysteresis loop of an arbitrary ferromagnetic material, showing the definitions of magnetic hysteresis parameters.

In an increasing field the magnetization climbs to a saturation value (equivalent to M_S), at which the individual magnetic moments are aligned with the applied field. If the magnetizing field is now reduced progressively to zero, the magnetization of a ferromagnetic material does not decrease along the path of increasing magnetization. It decreases along a new path such that in zero field the magnetization does not return to zero, but retains part of the induced magnetization. The residual magnetization is called the *remanence*, or *isothermal remanent magnetization* (IRM); if the sample was magnetized to saturation, the remanence is a saturation IRM (M_{rs}). For a given ferromagnetic mineral, the ratio M_{rs}/M_S depends on grain size. If a magnetic field is applied in the opposite direction to the IRM, it remagnetizes part of the material in the antiparallel direction. For a particular value B_c of the reverse field (called the *coercive force* or *coercivity*), the induced reverse magnetization exactly cancels the original remanence and the net magnetization is zero. If the reverse field is removed at this stage, the residual remanence is smaller than the original IRM. By repeating the process in ever stronger reverse fields a back-field B_{cr} (called the *coercivity of remanence*) is found which gives a reverse remanence that exactly cancels the IRM, so that the residual remanence is zero. The ratio B_{cr}/B_c also depends on grain size. Rock-forming magnetic minerals often have natural remanences with very high coercive properties.

When a ferromagnetic material is heated, its spontaneous magnetization disappears at the *ferromagnetic Curie temperature* (T_c). At temperatures higher than the *paramagnetic Curie temperature* (θ) the susceptibility k becomes the paramagnetic susceptibility and $1/k$ is proportional to $(T - \theta)$ as given by the Curie–Weiss law (Eq. 2). The paramagnetic Curie temperature for a ferromagnetic solid is several degrees higher than its ferromagnetic Curie temperature, T_c. The gradual transition from ferromagnetic to paramagnetic behavior is explained by persistence of the ferromagnetic molecular field due to short-range magnetic order above T_c.

Antiferromagnetism

In oxide crystals the oxygen ions usually keep the metal ions far apart, so that direct exchange of electrons between the metal ions is not possible. However, in certain minerals, interaction between magnetic spins becomes possible by the exchange of electrons from one metal ion to the other through the electron "cloud" of the oxygen ion. This *indirect exchange* (or *superexchange*) process results in antiparallel directions of adjacent atomic magnetic moments (Fig. C2b), giving two sublattices with equal and opposite intrinsic magnetic moments. As a result, the susceptibility of an antiferromagnetic crystal is weak and positive, and remanent magnetization is not possible. The antiferromagnetic alignment breaks down at the *Néel temperature*, above which paramagnetic behavior is shown. The Néel temperature T_N of many antiferromagnetic substances is lower than room temperature, at which they are paramagnetic. A common example of an antiferromagnetic mineral is ilmenite ($FeTiO_3$), which has a Néel temperature of 50 K.

Parasitic Ferromagnetism

When an antiferromagnetic crystal contains defects, vacancies, or impurities, some of the antiparallel spins are unpaired. A weak “defect moment” can arise due to these lattice imperfections. Also, if the spins are not exactly antiparallel but are inclined at a small angle, they do not cancel out completely and again a ferromagnetic type of magnetization can result (Fig. C2c). Materials that exhibit this form of *parasitic ferromagnetism* have the typical characteristics of a true ferromagnetic metal, including hysteresis, a spontaneous magnetization, and a Curie temperature. An important geological example is the common iron mineral hematite (α-Fe_2O_3), in which both the spin-canted and defect moments contribute to the ferromagnetic properties. Hematite has a variable, weak spontaneous magnetization of about 2000 A m^{-1}, very high coercivity, and a Curie temperature around 675°C. The variable magnetic properties are due to variation in the relative importances of the defect and spin-canted moments.

Ferrimagnetism

The metallic ions in an antiferromagnetic crystal occupy the voids between the oxygen ions. In certain crystal structures, of which the most important geological example is the spinel structure, the sites of the metal ions differ from each other in the coordination of the surrounding oxygen ions. Tetrahedral sites have four oxygen ions as nearest neighbors and octahedral sites have six. The tetrahedal and octahedral sites form two sublattices. In a normal spinel the tetrahedral sites are occupied by divalent ions and the octahedral sites by Fe^{3+} ions. The most common iron oxide minerals have an *inverse spinel* structure. Each sublattice has an equal number of Fe^{3+} ions. The same number of divalent ions (e.g., Fe^{2+}) occupy other octahedral sites, while the corresponding number of tetrahedral sites is empty.

When the indirect exchange process involves antiparallel and unequal magnetizations of the sublattices (Fig. C2d), resulting in a net spontaneous magnetization, the phenomenon is called *ferrimagnetism*. Ferrimagnetic materials (called *ferrites*) exhibit magnetic hysteresis and retain a remanent magnetization when they are removed from a magnetizing field. Above a given temperature – sometimes called the ferrimagnetic Néel temperature but more commonly the Curie temperature – the long-range molecular order breaks down and the mineral becomes paramagnetic. The most important ferrimagnetic mineral is magnetite (Fe_3O_4), but maghemite, pyrrhotite, and goethite are also significant contributors to the magnetic properties of rocks.

Appendix D

Magnetic Anisotropy

Magnetostatic (Shape) Anisotropy

In strongly magnetic materials the shape of the magnetized object causes a *magnetostatic* anisotropy. In rocks this effect is associated with the shapes of the individual grains of ferrimagnetic mineral in the rock, and to a lesser extent with the shape of the rock sample. The origin of the shape anisotropy can be conveniently explained with the aid of the concept of magnetic poles.

The spontaneous magnetization of a uniformly magnetized material can be pictured as giving rise to a distribution of poles on the free end surfaces. The field lines of the magnetic B-field form closed loops, whereas the H-field begins and ends on boundary surfaces, at which it is discontinuous. The field lines of the magnetic field H *outside* a magnet are parallel to the B-field and are directed from the surface distribution of N-poles to the surface distribution of S-poles. The H-field *inside* the magnet is also directed from the distribution of N-poles on one end to the S-poles on the other end. It forms a *demagnetizing field* (H_d) that opposes the magnetization. The strength of the demagnetizing field varies directly with the surface density of the magnetic pole distribution on the end surfaces of the magnet, and inversely with the distance between these surfaces. H_d thus depends on the shape of the magnet and the intensity of magnetization; it can be written as

$$\mathbf{H}_\mathrm{d} = -N\mathbf{M}.$$

(D1)

N is called the *demagnetizing factor*. It is a dimensionless constant determined by the shape of the magnetic grain. It can be computed for a geometrical shape, such as a triaxial ellipsoid. The demagnetizing factors N_1, N_2, and N_3 parallel to the symmetry axes of an ellipsoid satisfy the relationship

$$N_1 + N_2 + N_3 = 1. \tag{D2}$$

The magnetostatic energy of the interaction of the grain magnetization with the demagnetizing field is called the *demagnetizing* energy (E_d). For a grain with uniform magnetization of intensity M in a direction with demagnetizing factor N,

$$E_d = \frac{\mu_0}{2} N M^2. \tag{D3}$$

The magnetostatic anisotropy of a small grain shaped like a prolate ellipsoid is due to the variation of E_d with the shape of the ellipsoid. When the spontaneous magnetization M_s is along the long axis of the ellipsoid (Fig. D1a), the opposing pole distributions are further away from each other and their surface density is lower than when M_s is parallel to the short axis (Fig. D1b). The demagnetizing field and energy are lowest when M_s is parallel to the long axis, which is the energetically favored direction of magnetization. The demagnetizing energy is larger in any other direction. If N_1 is the demagnetizing factor for the long axis and N_2 that for the short axis ($N_1 < N_2$), the difference in energy between the two directions of magnetization defines a magnetostatic anisotropy energy density E_d given by

$$E_d = \frac{\mu_0}{2} (N_2 - N_1) M^2, \tag{D4}$$

which is minimum when M_s is parallel to the longest dimension of the grain.

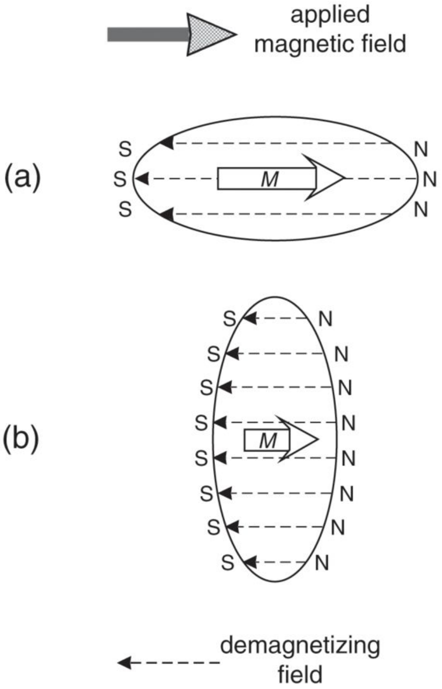

applied magnetic field
(a)
S
S
S
M
N
N
N
(b)
S
S
S
S
S
S
S
S
M
N
N
N
N
N
N
N
N
demagnetizing field

Fig. D1 The origin of shape anisotropy in a magnetized ellipsoid. The same applied field produces a stronger magnetization in (a) than in (b), because the internal demagnetizing field due to magnetic pole distributions on the end surfaces is weaker in (a) than in (b).

Shape-dependent magnetic anisotropy is important in minerals that have a high spontaneous magnetization. The more elongate the grain is, the higher the shape anisotropy will be. It is the predominant form of anisotropy in very fine grains of magnetite (and maghemite) if the longest axis exceeds the shortest axis by only about 20%.

Magnetocrystalline Anisotropy

The direction of the spontaneous magnetization (M_S) in a ferromagnetic metal is not arbitrary. The molecular field that produces M_s originates in the direct exchange of electron spins between neighboring atoms in a metal. The symmetry of the lattice structure of the metal affects the exchange process and gives rise to a *magnetocrystalline* anisotropy energy, which has a minimum value when M_s is parallel to a favored direction, referred to as the *easy axis* (or *easy direction*) of magnetization. The simplest form of magnetic anisotropy is uniaxial anisotropy, when a metal has only a single easy axis. For example, cobalt has a hexagonal structure and the easy direction is parallel to the *c*-axis at room temperature. Iron and nickel have cubic unit cells; at room temperature, the easy axes in iron are the edges of the cube, but the easy axes in nickel are the body diagonal directions.

Magnetocrystalline anisotropy is also exhibited by ferrites, including the geologically important ferrimagnetic minerals. The exchange process in a ferrite is indirect, but energetically preferred easy axes of magnetization arise that reflect the symmetry of the crystal structure. This gives rise to different forms of the anisotropy in hematite and magnetite.

Hematite has a rhombohedral or hexagonal structure and a *uniaxial anisotropy* with regard to the *c*-axis of symmetry. Oxygen ions form a close-packed hexagonal lattice in which two-thirds of the octahedral interstices are occupied by ferric (Fe^{3+}) ions. When the spontaneous magnetization makes an angle ϕ with the *c*-axis of the crystal, the uniaxial anisotropy energy density E_a can be written to first order as

$$E_a = K_u \sin^2\phi. \tag{D5}$$

K_u is called the uniaxial magnetocrystalline anisotropy constant. Its value in hematite is around -10^3 J m^{-3} at room temperature. The negative value of K_u in Eq. (D5) means that E_a decreases as the angle ϕ increases. It is minimum when $\sin 2\phi$ is maximum, i.e., when ϕ is 90°. As a result, the spontaneous magnetization lies in the basal plane of the hematite crystal at room temperature.

Because of its inverse spinel structure, magnetite has *cubic anisotropy*. Let the direction of the spontaneous magnetization be given by direction cosines α_1, α_2, and α_3 relative to the edges of the cubic unit cell (as in Box 2.2). The magnetocrystalline anisotropy energy density is then given by

$$E_a = K_1(\alpha_1^2\alpha_2^2 + \alpha_2^2\alpha_3^2 + \alpha_3^2\alpha_1^2) + K_2\alpha_1^2\alpha_2^2\alpha_3^2. \tag{D6}$$

In magnetite $K_1 = -1.36 \times 10^4$ J m^{-3} and $K_2 = -0.44 \times 10^4$ J m^{-3} at room temperature. Because the anisotropy constants are negative, the anisotropy energy density E_a is minimum when the spontaneous magnetization is along a [111] body diagonal, and this is the magnetocrystalline easy axis of magnetization at room temperature.

Magnetostrictive Anisotropy

The process of magnetizing some materials causes them to change shape. Within the crystal lattice the interaction energy between atomic magnetic moments depends on their separations (called the bond length) and on their orientations, i.e., on the direction of magnetization. If an applied field changes the orientations of the atomic magnetic moments so that the interaction energy is increased, the bond lengths adjust to reduce the total energy. This produces strains and changes the shape of the ferromagnetic specimen. The behavior is called *magnetostriction*. A material that elongates in the direction of magnetization exhibits positive magnetostriction, while a material that shortens parallel to the magnetization shows negative magnetostriction. The maximum difference in magnetostrictive strain, which occurs between the demagnetized state and that of saturation magnetization, is called the *saturation magnetostriction*, denoted λ_s.

The inverse effect is also possible. For example, if pressure is applied to one face of a cubic crystal, it shortens elastically along the direction of the applied stress and expands in directions perpendicular to it. These strains alter the separations of atomic magnetic moments, thereby perturbing the effects that give rise to magnetocrystalline anisotropy. Thus the application of stress to a magnetic material can change its magnetization. On a submicroscopic scale the stress field that surrounds a vacancy, defect, or dislocation in the crystal structure can locally affect the orientations of ionic magnetic spins.

Magnetostriction is a further source of anisotropy in magnetic minerals. The magnetostrictive anisotropy energy density E_a depends on the amount and direction of the stress σ. If the saturation magnetization makes an angle θ to the stress, E_a is given for a uniaxial magnetic mineral by

$$E_a = \frac{3}{2}\lambda_s \sigma \cos^2\theta. \tag{D7}$$

This is the simplest expression for magnetostrictive energy. It assumes that the magnetostriction is isotropic, i.e., that it has the same value in all directions. This condition is fulfilled if the magnetocrystalline axes of the ferrimagnetic minerals in a rock are randomly distributed. The magnetostrictive energy of a cubic mineral is more complicated. Instead of a single magnetostriction constant λ_s, separate constants λ_{100} and λ_{111} are required for the saturation magnetostriction along the [100] and [111] directions, respectively, corresponding to the edge and body diagonal directions of the cubic unit cell.

In magnetite, the magnetostrictive energy is more than an order of magnitude smaller than the magnetocrystalline energy at room temperature. Consequently, magnetostriction plays only a secondary role in determining the direction of magnetization of magnetite grains. However, in minerals that have high magnetostriction (e.g., titanomagnetites with a compositional factor $x \approx 0.65$, see Section 12.2.1) the magnetostrictive energy may be significant in determining easy directions of magnetization, and the magnetization may be sensitive to modification by deformation.

Bibliography

As well as references to works cited in the captions of figures, in the titles of tables, and in the text of the book, this list contains references to historically important works mentioned in the text.

Achache, J., Le Mouël, J. L., and Courtillot, V. 1981. Long-period variations and mantle conductivity: an inversion using Bailey's method. *Geophys. J. R. Astr. Soc.*, **65**, 579–601.

Allègre, C. J., Manhès, G., and Göpel, C. 1995. The age of the Earth. *Geochim. Cosmochim. Acta*, **59**, 1445–1456.

Alvarez, L. W., Alvarez, W., Asaro, F., and Michel, H. V. 1980. Extraterrestrial cause for the Cretaceous–Tertiary extinction. *Science*, **208**, 1095–1108.

Anzellini, S., Dewaele, A., Mezouar, M., Loubeyre, P., and Morard, G. 2013. Melting of iron at Earth's inner core boundary based on fast X-ray diffraction. *Science*, **340**, 464–466.

Atwater, T. 1970. Implications of plate tectonics for the Cenozoic tectonic evolution of western North America. *Geol. Soc. Amer. Bull.*, **81**, 3513–3536.

Banks, R. J. 1969. Geomagnetic variations and the electrical conductivity of the upper mantle. *Geophys. J. R. Astr. Soc.*, **17**, 457–487.

Barazangi, M. and Dorman, J. 1969. World seismicity maps compiled from ESSA, Coast and Geodetic Survey, epicenter data, 1961–1967. *Bull. Seism. Soc. Amer.*, **59**, 369–380.

Barraclough, D. R. 1974. Spherical harmonic analyses of the geomagnetic field for eight epochs between 1600 and 1910. *Geophys. J. R. astr. Soc.*, **36**, 497–513.

Barton, C. E. 1989. Geomagnetic secular variation: direction and intensity. In *The Encyclopedia of Solid Earth Sciences*, James, D. E., ed., New York: Van Nostrand Reinhold, pp. 560–577.

Becquerel, H. 1896. Sur les radiations invisibles émises par les corps phosphorescents. *Comptes Rendues de l'Académie des Sciences*, **122**, 501.

Benioff, H. 1954. Orogenesis and deep crustal structure: additional evidence from seismology. *Bull. Geol. Soc. Amer.*, **65**, 385–400.

Berger, A. and Loutre, M. F. 2004. Astronomical theory of climate change. *J. Phys. IV France*, **121**, 1–35.

Bird, P. 2003. An updated digital model of plate boundaries. *Geochem., Geophys., Geosystems*, **4**

Bloxham, J., Zatman, S. A., and Dumberry, M. 2002. The origin of geomagnetic jerks. *Nature*, **420**, 66–68.

Bodell, J. M. and Chapman, D. S. 1982. Heat flow in the north-central Colorado Plateau. *J. Geophys. Res.*, **87**, 2869–2884.

Bott, M. H. P. 1982. *The Interior of the Earth* (2nd ed.), London: Edward Arnold.

Bullard, E. C., Everett, J. E., and Smith, A. G. 1965. A symposium on continental drift, IV: the fit of the continents around the Atlantic. *Phil. Trans. Roy. Soc. London, Ser. A*, **258**, 41–51.

Busse, F. H. 1989. Fundamentals of thermal convection. In *Mantle Convection: Plate Tectonics and Global Dynamics*, Peltier, W. R., ed., New York: Gordon and Breach, pp. 23–95.

Butler, R. F. 1992. *Paleomagnetism: Magnetic Domains to Geologic*

Terranes, Boston, MA: Blackwell Scientific.

Caldwell, J. G. and Turcotte, D. L. 1979. Dependence of the thickness of the elastic oceanic lithosphere on age. *J. Geophys. Res.*, **84**, 7572–7576.

Cande, S. C. and Kent, D. V. 1992. A new geomagnetic polarity time scale for the Late Cretaceous and Cenozoic. *J. Geophys. Res.*, **97**, 13917–13951.

Cande, S. C. and Kent, D. V. 1995. Revised calibration of the geomagnetic polarity timescale for the Late Cretaceous and Cenozoic. *J. Geophys. Res.*, **100**, 6093–6095.

Carey, S. W. 1958. A tectonic approach to continental drift. In *Continental Drift, a Symposium*, Carey, S. W., ed., Hobart: University Tasmania, pp. 172–355.

Carter, W. E. 1989. Earth orientation. In *The Encyclopedia of Solid Earth Geophysics*, James, D. E., ed., New York: Van Nostrand Reinhold, pp. 231–239.

Cathles, L. M. 1975. *The Viscosity of the Earth's Mantle*, Princeton, NJ: Princeton University Press.

Channell, J. E. T., Erba, E., Nakanishi, M., and Tamaki, K. 1995. Late Jurassic–Early Cretaceous time scales and oceanic magnetic anomaly block models. In *Geochronology, Timescales, and Stratigraphic Correlation*, Berggren, W. A., Kent, D. V., Aubry, M., and Hardenbol, J., ed., Tulsa, OK: Society for Sedimentary Geology (SEPM), pp. 51–64.

Chapman, S. and Bartels, J. 1940. *Geomagnetism, Vol. I: Geomagnetic and Related Phenomena, Vol. II: Analysis of Data and Physical Theories*, Oxford: Oxford University Press.

Chapple, W. M. and Tullis, T. E. 1977. Evaluation of the forces that drive plates. *J. Geophys. Res.*, **82**, 1967–1984.

Chevallier, R. 1925. L'aimantation des lavas de l'Etna et l'orientation du champs terrestre en Sicile du 12e au 17e siècle. *Ann. Phys. Ser. 10*, **4**, 5–162.

Christensen, A.V., Auken, E., and Sørensen, K. 2009. The transient electromagnetic method. In: *Groundwater Geophysics*, R. Kirsch, ed., Berlin: Springer, pp. 179–226.

Connerney, J. E. P., Kotsiaros, S., Oliversen, R. J., et al. 2018. A new model of Jupiter's magnetic field from Juno's first nine orbits. *Geophys. Res. Lett.*, **45**, 2590–2596.

Conrad, V. 1925. Laufzeitkurven des Tauernbebens vom 28. November, 1923. *Mitt. ErdbKomm. Wien*, **59**, 1–23.

Cornell, C. A., Banon, H., and Shakal, A. F. 1979. Seismic motion and response prediction alternatives. *Earthq. Eng. Struct. Dyn.*, **7**, 295–315.

Courtillot, V. and Besse, J. 1987. Magnetic field reversals, polar wander, and core–mantle coupling. *Science*, **237**, 1140–1147.

Cox, A. and Hart, R. B. 1986. *Plate Tectonics*, Boston, MA: Blackwell Scientific.

Cox, A. V. 1982. Magnetostratigraphic time scale. In *A Geologic Time Scale*, Harland, W. B., Cox, A. V., Llewellyn, P. G., et al., ed., Cambridge: Cambridge University Press, pp. 63–84.

Creer, K. M. and Irving, E. 2012. Testing continental drift: constructing the first palaeomagnetic path of polar wander (1954). *Earth Sci. History*, **31**, 111–145.

Crough, S. T. and Jurdy, D. M. 1980. Subducted lithosphere, hotspots, and the geoid. *Earth Planet. Sci. Lett.*, **48**, 15–22.

Dalrymple, G. B. 1991. *The Age of the Earth*, Stanford, CA: Stanford University Press.

Dalrymple, G. B., Clague, D. A., and Lanphere, M. A. 1977. Revised age for Midway volcano, Hawaiian volcanic chain. *Earth Planet. Sci. Lett.*, **37**, 107–116.

Davies, J. H. and Davies, D. R. 2010. Earth's surface heat flux. *Solid Earth*, **1**, 5–24.

Davis, J. C. 1986. *Statistics and Data Analysis in Geology*, New York: Wiley.

DeMets, C., Gordon, R. G., and Argus, D. F. 2010. Geologically current plate motions. *Geophys. J. Int.*, **181**, 1–80.

DeMets, C., Gordon, R. G., Argus, D. F., and Stein, S. 1990. Current plate motions. *Geophys. J. Int.*, **101**, 425–478.

Dietz, R. S. 1961. Continent and ocean basin evolution by spreading of the sea floor. *Nature*, **190**, 854–857.

Dobrin, M. B. 1976. *Introduction to Geophysical Prospecting* (3rd ed.), New York: McGraw-Hill.

Dobrin, M. B. and Savit, C. H. 1988. *Introduction to Geophysical Prospecting* (4th ed.), New York: McGraw-Hill.

Domeier, M. and Torsvik, T. H. 2014. Plate tectonics in the Late Paleozoic. *Geoscience Frontiers*, **5**, 303–350.

du Toit, A. L. 1937. *Our Wandering Continents*, Edinburgh: Oliver and Boyd.

Dziewonski, A. M. 1989. Earth structure, global. In *The Encyclopedia of Solid Earth Geophysics*, James, D. E., ed., New York: Van Nostrand Reinhold, pp. 331–359.

Dziewonski, A. M. and Anderson, D. L. 1981. Preliminary Reference Earth Model (PREM). *Phys. Earth Planet. Inter.*, **25**, 297–356.

Engeln, J. F., Wiens, D. A., and Stein, S. 1986. Mechanisms and depths of Atlantic transform earthquakes. *J. Geophys. Res.*, **91**, 548–577.

European Seismological Commission. 1998. *European Macroseismic Scale 1998*, Luxembourg: Centre Européen de Géodynamique et de Séismologie, Conseil de l'Europe.

Evans, M. E. and Heller, F. 2003. *Environmental Magnetism: Principles and Applications of Enviromagnetics*, New York: Academic Press.

Evans, M. E. and McElhinny, M. W. 1969. An investigation of the origin of stable remanence in magnetite-bearing igneous rocks. *J. Geomag. Geoelect.*, **21**, 757–773.

Finlay, C. C. and Jackson, A. 2003. Equatorially dominated magnetic field change at the surface of Earth's core. *Science,* **300**, 2084–2086.

Fisher, R. A. 1953. Dispersion on a sphere. *Proc. R. Soc. Lond., Ser. A*, **217**, 295–305.

Forsyth, D. and Uyeda, S. 1975. On the relative importance of the driving forces of plate motion. *Geophys. J. R. Astr. Soc.*, **43**, 163–200.

Frank, M., Schwarz, B., Baumann, S., et al. 1997. A 200 kyr record of cosmogenic radionuclide production rate and geomagnetic field intensity from 10-Be in globally stacked deep-sea sediments. *Earth Planet. Sci. Lett.*, **149**, 121–129.

Gambis, D. and Bizouard, C. 2014. *IERS Annual Report 2013,* Frankfurt am Main: Bundesamt für Kartographie und Geodäsie.

Giardini, D., Woessner, J., and Danciu, L. 2014. Mapping Europe's seismic hazard. *EOS,* **95**, 261–262.

Glatzmaier, G. A. and Roberts, P. H. 1995. A three-dimensional convective dynamo solution with rotating and conducting inner core and mantle. *Phys. Earth Planet. Inter.*, **91**, 63–75.

Gradstein, F. M., Ogg, J. G., Schmitz, M. D., and Ogg, G. M. (ed.) 2012. *The Geologic Time Scale 2012,* Boston, MA: Elsevier B.V.

Grayver, A. V., Munch, F. D., Kuvshinov, A., et al. 2017. Joint inversion of satellite-detected tidal and magnetospheric signals constrains electrical conductivity and water content of the upper mantle and transition zone. *Geophys. Res. Lett.*, **44**, 6074–6081.

Grow, J. A. and Bowin, C. O. 1975. Evidence for high-density crust and mantle beneath the Chile Trench due to the descending lithosphere. *J. Geophys. Res.*, **80**, 1449–1458.

Gubler, E. 1991. Recent crustal movements in Switzerland: vertical movements. In *Report on the Geodetic Activities in the Years 1987 to 1991: 5.* Geodynamics, Kahle, H. G. and Jeanrichard, F., ed., Zürich: Swiss Geodetic Commission and Federal Office of Topography, pp. 36 and Map 35.

Gutenberg, B. 1914. Über Erdbebenwellen. VIIA. Beobachtungen an Registrierungen von Fernbeben in Göttingen und Folgerungen über die Konstitution des Erdkörpers. *Nachr. Ges. Wiss. Göttingen, Math.-Phys. Klasse*, 1–52 and 125–176.

Gutenberg, B. 1945. Amplitudes of surface waves and magnitudes of shallow earthquakes. *Bull. Seism. Soc. Am.*, **35**, 3–12.

Gutenberg, B. 1956. The energy of earthquakes. *Quart. J. Geol. Soc. Lond.*, **112**, 1–14.

Gutenberg, B. and Richter, C. F. 1954. *The Seismicity of the Earth and Associated Phenomena*, Princeton, NJ: Princeton University Press.

Guyodo, Y. and Valet, J.-P. 1996. Relative variations in geomagnetic intensity from sedimentary records: the past 200,000 years. *Earth Planet. Sci. Lett.*, **143**, 23–36.

Halliday, A. N. 2001. In the beginning …. *Nature*, **409**, 144–145.

Harland, W. B., Armstrong, R. L., Cox, A. V., et al. 1990. *A Geologic Time Scale 1989*, Cambridge: Cambridge University Press.

Hasegawa, A. 1989. Seismicity: subduction zone. In *The Encyclopedia of Solid Earth Geophysics*, James, D. E., ed., New York: Van Nostrand

Reinhold, pp. 1054–1061.

Hasterok, D., Chapman, D. S., and Davis, E. E. 2011. Oceanic heat flow: implications for global heat loss. *Earth Planet. Sci. Lett.*, **311**, 386–395.

Hauck, C. and Vonder Mühll, D. 2003. Inversion and interpretation of two-dimensional geoelectrical measurements for detecting permafrost in mountainous regions. *Permafrost and Periglacial Processes*, **14**, 305–318.

Heirtzler, J. R., Dickson, G. O., Herron, E. M., Pitman III, W. C. and Le Pichon, X. 1968. Marine magnetic anomalies, geomagnetic field reversals, and motions of the ocean floor and continents. *J. Geophys. Res.*, **73**, 2119–2136.

Heirtzler, J. R., Le Pichon, X., and Baron, J. G. 1966. Magnetic anomalies over the Reykjanes Ridge. *Deep Sea Res.*, **13**, 427–443.

Hess, H. H. 1962. History of ocean basins. In *Petrologic Studies: a Volume in Honor of A. F. Buddington*, Engel, A. E., James, H. L., and Leonard, B. F., ed., Boulder, CO: Geological Society of America, pp. 599–620.

Holliger, K. and Kissling, E. 1992. Gravity interpretation of a unified 2-D acoustic image of the central Alpine collision zone. *Geophys. J. Int.*, **111**, 213–225.

Holmes, A. 1965. *Principles of Physical Geology* (2nd ed.), London; Thomas Nelson and Sons.

Hood, L. L., Mitchell, D. L., Lin, R. P., Acuna, M. H. and Binder, A. B. 1999. Initial measurements of the lunar induced magnetic dipole moment using Lunar Prospector magnetometer data. *Geophys. Res. Lett.*, **26**, 2327–2330.

Huang, P. Y., Solomon, S. C., Bergman, E. A., and Nabelek, J. L. 1986. Focal depths and mechanisms of Mid-Atlantic Ridge earthquakes from body waveform inversion. *J. Geophys. Res.*, **91**, 579–598.

Hurst, R. W., Bridgwater, D., Collerson, K. D., and Weatherill, G. W. 1975. 3600 m.y. Rb-Sr ages from very early Archean gneisses from Saglek Bay, Labrador. *Earth Planet. Sci. Lett.*, **27**, 393–403.

Irving, E. 1957. Rock magnetism: a new approach to some palaeogeographic problems. *Adv. Phys.*, **6**, 194–218.

Irving, E. 1977. Drift of the major continental blocks since the Devonian. *Nature,* **270**, 304–309.

Irving, E. and Major, A. 1964. Post-depositional detrital remanent magnetization in artificial and natural sediments. *Sedimentology*, **3**, 135–143.

Isacks, B. and Molnar, P. 1969. Mantle earthquake mechanisms and the sinking of the lithosphere. *Nature,* **223**, 1121–1124.

Isacks, B. L. 1989. Seismicity and plate tectonics. In *The Encyclopedia of Solid Earth Geophysics*, James, D. E., ed., New York: Van Nostrand Reinhold, pp. 1061–1071.

Jackson, A., Jonkers, A. R. T., and Walker, M. R. 2000. Four centuries of geomagnetic secular variation from historical records. *Phil. Trans. R. Soc. London A,* **358**, 957–990.

Jarvis, G. T. and Peltier, W. R. 1989. Convection models and geophysical observations. In *Mantle Convection: Plate Tectonics and Global Dynamics*, Peltier, W. R., ed., New York: Gordon and Breach, pp. 479–593.

Johnson, C. L., Purucker, M. E., Korth, H., et al. 2012. MESSENGER observations of Mercury’s magnetic field structure. *J. Geophys. Res.*, **117**, E00L14.

Johnson, H. P. and Carlson, R. L. 1992. Variation of sea floor depth with age: a test of models based on drilling results. *Geophys. Res. Lett.*, **19**, 1971–1974.

Joly, J. 1899. An estimation of the geological age of the Earth. *Ann. Rept. Smithson. Inst.*, **1899**, 247–288.

Jouzel, J., Koster, R. D., Zuozzo, R. J., and Russell, G. L. 1994. Stable water isotope behavior during the last glacial maximum: a general circulation model analysis. *J. Geophys. Res.*, **99**, 25, 791–725, 801.

Kanamori, H. and Brodsky, E. E. 2004. The physics of earthquakes. *Rep. Prog. Phys.*, **67**, 1429–1496.

Karnik, V. 1969. *Seismicity of the European Area, Part 1*, Dordrecht: D. Reidel and Company.

Keen, C. E. and Tramontini, C. 1970. A seismic refraction survey on the Mid-Atlantic Ridge. *Geophys. J. R. Astr. Soc.*, **20**, 473–491.

Kennett, B. L. N. and Engdahl, E. R. 1991. Traveltimes for global earthquake location and phase identification. *Geophys. J. Int.*, **105**, 429–465.

Kennett, B. L. N., Engdahl, E. R., and Buland, R. 1995. Constraints on seismic velocities in the Earth from traveltimes. *Geophys. J. Int.*, **122**, 108–124.

Kissling, E. 1993. Seismische Tomographie: Erdbebenwellen durchleuchten unseren Planeten. *Zürich: Vierteljahresschrift Naturforsch. Ges.* **138**, 1–20.

Klingelé, E. and Kissling, E. 1982. Zum Konzept der isostatischen Modelle in Gebirgen am Beispiel der Schweizer Alpen, Geodätisch-Geophysikalische Arbeiten in der Schweiz. *Schweiz. Geodät. Komm.*, **35**, 3–36.

Klingelé, E. and Olivier, R. 1980. Die neue Schwere-Karte der Schweiz (Bouguer-Anomalien). *Beitr. Geologie der Schweiz, Serie Geophys.*, **20**, 1–93.

Koelemeijer, P., Ritsema, J., Deuss, A., and van Heijst, H.-J. 2016.

SP12RTS: a degree-12 model of shear- and compressional-wave velocity for the Earth’s mantle. *Geophys. J. Int.*, **204**, 1024–1039.

Köppen, W. and Wegener, A. 1924. *Die Klimate der geologischen Vorzeit*, Berlin: Bornträger.

Korte, M. and Constable, C. G. 2005. The geomagnetic dipole moment over the last 7000 years: new results from a global model. *Earth Planet. Sci. Lett.*, **236**, 348–358.

Kreemer, C., Blewitt, G., and Klein, E. C. 2014. A geodetic plate motion and Global Strain Rate Model. *Geochem., Geophys, Geosyst.*, **15**, 3849–3889.

Kurtz, R. D., DeLaurier, J. M., and Gupta, J. C. 1986. A magnetotelluric sounding across Vancouver Island detects the subducting Juan de Fuca plate. *Nature*, **321**, 596–599.

Lambeck, K., Purcell, A., and Zhao, S. 2017. The North American late Wisconsin ice sheet and mantle viscosity from glacial rebound analyses. *Quat. Sci. Rev.*, **158**, 172–210.

Lay, T., Kanamori, H., Ammon, C. J., et al. 2005. The Great Sumatra–Andaman Earthquake of 26 December 2004. *Science*, **308**, 1127–1133.

Lay, T. and Wallace, T. C. 1995. *Modern Global Seismology*, San Diego, CA: Academic Press.

Le Pichon, X. 1968. Sea-floor spreading and continental drift. *J. Geophys. Res.*, **73**, 3661–3697.

Lehmann, I. 1936. P′. *Bur. Centr. seism. Internat. A*, **14**, 3–31.

Lerch, F. J., Klosko, S. M., Laubscher, R. E., and Wagner, C. A. 1979. Gravity model improvement using Geos 3 (GEM 8 and 10). *J. Geophys. Res.*, **84**, 3897–3916.

Lindner, H., Militzer, H., Rösler, R., and Scheibe, R. 1984. Bearbeitung und Interpretation der gravimetrischen und magnetischen

Messergebnisse. In *Angewandte Geophysik, I: Gravimetrie und Magnetik*, Militzer, H. and Weber, F., ed., Vienna: Springer-Verlag, pp. 226–283.

Love, A. E. H. 1911. *Some Problems of Geodynamics*, Cambridge: Cambridge University Press.

Lowrie, W. and Alvarez, W. 1975. Paleomagnetic evidence for rotation of the Italian peninsula. *J. Geophys. Res.*, **80**, 1579–1592.

Lowrie, W. and Alvarez, W. 1977. Upper Cretaceous–Paleocene magnetic stratigraphy at Gubbio, Italy: III. Upper Cretaceous magnetic stratigraphy. *Geol. Soc. Amer. Bull.*, **88**, 374–377.

Ludwig, W. J., Nafe, J. E., and Drake, C. L. 1970. Seismic refraction. In *The Sea, Vol. 4*, Maxwell, A. E., ed, New York: Wiley-Interscience, pp. 53–84.

Luzum, B., Capitaine, N., Fienga, A., et al. 2011. The IAU 2009 system of astronomical constants: the report of the IAU working group on Numerical Standards for Fundamental Astronomy. *Celestial Mechanics and Dynamical Astronomy*, **110**, 293–304.

MacDonald, K. L. 1957. Penetration of the geomagnetic secular field through a mantle with variable conductivity. *J. Geophys. Res.*, **62,** 117–141.

Marsh, J. G., Koblinsky, C. J., Zwally, H. J., Brenner, A. C., and Beckley, B. D. 1992. A global mean sea surface based upon GEOS 3 and Seasat altimeter data. *J. Geophys. Res.*, **97**, 4915–4921.

Massonnet, D. 1997. Satellite radar interferometry. *Scientific American*, **276**, 46–53.

Maurer, H. and Ansorge, J. 1992. Crustal structure beneath the northern margin of the Swiss Alps. *Tectonophysics*, **207**, 165–181.

Maus, S. 2008. The geomagnetic power spectrum. *Geophys. J. Int.*, **174**, 135–142.

Maus, S., Silva, L., and Hulot, G. 2008. Can core–surface flow models be used to improve the forecast of the Earth's main magnetic field? *J. Geophys. Res.*, **113**, B08102.

Mayer Rosa, D. and Müller, S. 1979. Studies of seismicity and selected focal mechanisms of Switzerland. *Schweiz. Miner. Petr. Mitt.*, **59**, 127–132.

McClusky, S., Balassanian, S., Barka, A., et al. 2000. Global Positioning System constraints on plate kinematics and dynamics in the eastern Mediterranean and Caucasus. *J. Geophys. Res.*, **105**, 5695–5719.

McElhinny, M. W. 1973. *Paleomagnetism and Plate Tectonics*, Cambridge: Cambridge University Press.

McKenzie, D. P. and Morgan, W. J. 1969. Evolution of triple junctions. *Nature*, **224**, 125–133.

McKenzie, D. P. and Parker, R. L. 1967. The North Pacific: an example of tectonics on a sphere. *Nature*, **216**, 1276–1280.

Meier, U., Trampert, J., and Curtis, A. 2009. Global variations of temperature and water content in the mantle transition zone from higher-mode surface waves. *Earth Planet. Sci. Lett.*, **282**, 91–101.

Militzer, H., Scheibe, R., and Seiberl, W. 1984. Angewandte Magnetik. In *Angewandte Geophysik, I: Gravimetrie und Magnetik*, Militzer, H. and Weber, F., ed., Vienna: Springer-Verlag, pp. 127–189.

Mohorovičić, A. 1909. Das Beben vom 8.x.1909. *Jb. met. Obs. Zagreb (Agram)*, **9**, 1–63.

Mohr, P. J., Newell, D. B., and Taylor, B. N. 2016. CODATA recommended values of the fundamental physical constants: 2014. *Rev. Mod. Phys.*, **88**, 1–71.

Molnar, P. 1988. Continental tectonics in the aftermath of plate tectonics. *Nature*, **335**, 131–137.

Morgan, W. J. 1968. Rises, trenches, great faults, and crustal blocks. *J. Geophys. Res.*, **73**, 1959–1982.

Morgan, W. J. 1982. Hotspot tracks and the opening of the Atlantic and Indian Oceans. In *The Sea, Vol. 7: The Oceanic Lithosphere*, Emiliani, C., ed., New York: Wiley-Intersciences, pp. 443–487.

Mueller, S. 1977. A new model of the continental crust. In *The Earth's Crust*, Heacock, J. G., ed., Washington, DC: American Geophysical Union, pp. 289–317.

Nagata, T. 1961. *Rock Magnetism*, Tokyo: Maruzen.

National Imaging and Mapping Agency. 2000. *Department of Defense World Geodetic System 1984*. Technical Report 8350.2.

Ni, J. and Barazangi, M. 1984. Seismotectonics of the Himalayan collision zone: geometry of the underthrusting Indian plate beneath the Himalaya. *J. Geophys. Res.*, **89**, 1147–1163.

Nomura, R., Hirose, K., Uesugi, K., et al. 2014. Low core–mantle boundary temperature inferred from the solidus of pyrolite. *Science*, **343**, 522–525.

Nuttli, O. W. 1973. The Mississippi Valley earthquakes of 1811 and 1812: intensities, ground motion and magnitudes. *Bull. Seism. Soc. Amer.*, **63**, 227–248.

Oldham, R. D. 1906. The constitution of the interior of the Earth, as revealed by earthquakes. *Q. Jl. Geol. Soc. London*, **62**, 456–475.

Ondoh, T. and Maeda, H. 1962. Geomagnetic storm correlation between the northern and southern hemisphere. *J. Geomag. Geoelectr.*, **14**, 22–32.

Opdyke, N. D. 1972. Paleomagnetism of deep-sea cores. *Rev. Geophys.*, **10**, 213–249.

Opdyke, N. D., Glass, B., Hays, J. D., and Foster, J. 1968.

Paleomagnetic study of Antarctic deep-sea cores. *Science*, **154**, 349–357.

Park, J., Song, T. A., Tromp, J., et al. 2005. Earth's free oscillations excited by the 26 December 2004 Sumatra–Andaman earthquake. *Science*, **308**, 1139–1144.

Parker, R. L. and Oldenburg, D. W. 1973. Thermal models of mid-ocean ridges. *Nature Phys. Sci.*, **242**, 137–139.

Parsons, B. and McKenzie, D. 1978. Mantle convection and the thermal structure of the plates. *J. Geophys. Res.*, **83**, 4485–4496.

Parsons, B. and Sclater, J. G. 1977. An analysis of the variation of ocean floor bathymetry and heat flow with age. *J. Geophys. Res.*, **82**, 803–827.

Pavlis, N. K., Holmes, S. A., Kenyon, S. C., and Factor, J. K. 2012. The development and evaluation of the Earth Gravitational Model 2008 (EGM2008). *J. Geophys. Res.*, **117**, B04406.

Pavoni, N. 1977. Erdbeben im Gebiet der Schweiz. *Eclogae geol. Helv.*, **70**, 2, 351–371.

Peltier, W. R. 1989. Mantle viscosity. In *Mantle Convection: Plate Tectonics and Global Dynamics*, Peltier, W. R., ed., New York: Gordon and Breach, pp. 389–478.

Peltier, W. R. 2004. Global glacial isostasy and the surface of the ice-age Earth: the ICE-5 G (VM2) model and GRACE. *Annu. Rev. Earth Planet. Sci.*, **32**, 111–149.

Peltier, W. R., Jarvis, G. T., Forte, A. M., and Solheim, L. P. 1989. The radial structure of the mantle general circulation. In *Mantle Convection: Plate Tectonics and Global Dynamics*, Peltier, W. R., ed., New York: Gordon and Breach, pp. 765–815.

Perosanz, F., Biancale, R., Loyer, S., et al. 2003. On board evaluation of the STAR accelerometer. In *First CHAMP Mission Results for Gravity,* *Magnetic and Atmospheric Studies*, Reigber, C., Lühr, H., and Schwintzer, P., ed., Heidelberg: Springer-Verlag, pp. 563.

Pitman III, W. C. and Heirtzler, J. R. 1966. Magnetic anomalies over the Pacific-Antarctic ridge. *Science,* **154**, 1164–1171.

Pitman III, W. C. and Talwani, M. 1972. Sea-floor spreading in the North Atlantic. *Geol. Soc. Amer. Bull.*, **83**, 619–646.

Pollack, H. N. and Huang, S. 2000. Climate reconstruction from subsurface temperatures. *Annu. Rev. Earth Planet. Sci.*, **28**, 339–365.

Pollack, H. N., Hurter, S. J., and Johnson, J. R. 1993. Heat flow from the Earth's interior: analysis of the global data set. *Rev. Geophys.*, **31**, 267–280.

Poutanen, M. and Steffen, H. 2014. land uplift at Kvarken Archipelago/High Coast UNESCO World Heritage area. *Geophysica,* **50**, 25–40.

Powell, W. G., Chapman, D. S., Balling, N., and Beck, A. E. 1988. Continental heat-flow density. In *Handbook of Terrestrial Heat-Flow Density Determinations*, Haenel, R., Rybach, L., and Stegena, L., ed., Dordrecht: Kluwer Academic Publishers, pp. 167–222.

Pozzo, M., Davies, C., Gubbins, D., and Alfè, D. 2012. Thermal and electrical conductivity of iron at Earth's core conditions. *Nature,* **485**, 355–358.

Prévot, M., Mankinen, E. A., Grommé, C. S., and Coe, R. S. 1985. How the geomagnetic field vector reverses polarity. *Nature,* **316**, 230–234.

Püthe, C., Kuvshinov, A., Khan, A., and Olsen, N. 2015. A new model of Earth's radial conductivity structure derived from over 10 yr of satellite and observatory magnetic data. *Geophys. J. Int.*, **203**, 1864–1872.

Reid, H. F. 1906. The elastic-rebound theory of earthquakes. *Bull. Dep. Geol. Univ. Calif.*, **6**, 413–444.

Ricard, Y. 2009. Physics of mantle convection. In *Treatise on Geophysics, vol. 7*, Bercovici, D., ed. Amsterdam: Elsevier, pp. 32–82.

Richards, P. G. 1989. Seismic monitoring of nuclear explosions. In *The Encyclopedia of Solid Earth Geophysics*, James, D. E., ed., New York: Van Nostrand Reinhold, pp. 1071–1089.

Richards, T. C. 1961. Motion of the ground on arrival of reflected longitudinal and transverse waves at wide-angle reflection distances. *Geophysics*, **26**, 277–297.

Richter, C. F. 1935. An instrumental earthquake magnitude scale. *Bull. Seismol. Soc. Amer.*, **25**, 1–32.

Roberts, A. P., Pike, C.R., and Verosub, K.L. 2000. First-order reversal curve diagrams: a new tool for characterizing the magnetic properties of natural samples. *J. Geophys. Res.*, **105**, 28461–28475.

Robinson, E. S. and Çoruh, C. 1988. *Basic Exploration Geophysics*, New York: Wiley.

Roy, R. F., Blackwell, D. D., and Birch, F. 1968. Heat generation of plutonic rocks and continental heat flow provinces. *Earth Planet. Sci. Lett.*, **5**, 1–12.

Runcorn, S. K. 1956. Paleomagnetic comparisons between Europe and North America. *Proc. Geol. Assoc. Canada*, **8**, 77–85.

Rutherford, E. and Soddy, F. 1902. The cause and nature of radioactivity. *Phil. Mag.*, **4**, 370–396 and 569–585.

Ryan, J. W., Clark, T. A., Ma, C., et al. 1993. Global scale tectonic plate motions with CDP VLBI data. In *Contributions to Space Geodesy: Crustal Dynamics*, Smith, D. E. and Turcotte, D. L., ed., Washington, DC: American Geophysical Union, pp. 37–50.

Rybach, L. 1976. Radioactive heat production in rocks and its relation to other petrophysical parameters. *Pure Appl. Geophys.*, **114**, 309–318.

Rybach, L. 1988. Determination of heat production rate. In *Handbook of Terrestrial Heat Flow Density Determination*, Haenel, R., Rybach, L., and Stegena, L., ed., Dordrecht: Kluwer Academic Publishers, p. 486.

Schaeffer, A. J. and Lebedev, S. 2013. Global shear speed structure of the upper mantle and transition zone. *Geophys. J. Int.*, **194**, 417–449.

Schneider, D. A. and Kent, D. V. 1990. The time average paleomagnetic field. *Rev. Geophys.*, **28**, 71–96.

Schubert, G., Yuen, D. A., and Turcotte, D. L. 1975. Role of phase transitions in a dynamic mantle. *Geophys. J. R. Astr. Soc.*, **42**, 705–735.

Sclater, J. C., Jaupart, C., and Galson, D. 1980. The heat flow through oceanic and continental crust and the heat loss of the earth. *Rev. Geophys. Space Phys.*, **18**, 269–311.

Sclater, J. C., Parsons, B., and Jaupart, C. 1981. Oceans and continents: similarities and differences in the mechanisms of heat loss. *J. Geophys. Res.*, **86**, 11535–11552.

Scotese, C. R., Gahagan, L. M., and Larson, R. L. 1988. Plate tectonic reconstructions of the Cretaceous and Cenozoic ocean basins. *Tectonophysics*, **155**, 27–48.

Sella, G. F., Dixon, T. H., and Mao, A. 2002. REVEL: a model of recent plate velocities from space geodesy. *J. Geophys. Res.*, **107**

Serson, P. H. 1973. Instrumentation for induction studies on land. *Phys. Earth Planet. Int.*, **7**, 313–322.

Singh, S. K., Dominguez, T., Castro, R., and Rodriguez, M. 1984. P waveform of large shallow earthquakes along the Mexico subduction zone. *Bull. Seism. Soc. Amer.*, **74**, 2135–2156.

Slack, H. A., Lynch, V. M., and Langan, L. 1967. The geomagnetic elements. *Geophysics*, **32**, 877–892.

Smith, A. G. and Hallam, A. 1970. The fit of the southern continents. *Nature*, **225**, 139–144.

Stacey, F. D. 1992. *Physics of the Earth*, Brisbane: Brookfield Press.

Stacey, F. D. and Davis, P. M. 2008. *Physics of the Earth*, Cambridge: Cambridge University Press.

Stegman, D. R., Jellinek, A. M., Zatman, S. A., Baumgardner, J. R., and Richards, M. A. 2003. An early lunar core dynamo driven by thermochemical mantle convection. *Nature*, **421**, 143–146.

Stein, C. and Stein, S. 1992. A model for the global variation in oceanic depth and heat flow with lithospheric age. *Nature*, **359**, 123–129.

Stephenson, F. R. and Morrison, L. V. 1984. Long-term changes in the rotation of the earth: 700 B.C. to A.D. 1980. *Phil. Trans. R. Soc. London, Ser. A*, **313**, 47–70.

Strahler, A. N. 1963. *The Earth Sciences*, New York: Harper and Row.

Swisher, C. C., Grajales-Nishimura, J. M., Montanari, A., et al. 1992. Coeval ^{40}Ar/^{39}Ar ages of 65.0 million years ago from Chicxulub crater melt rock and Cretaceous–Tertiary boundary tektites. *Science*, **257**, 954–958.

Sykes, L. R. 1967. Mechanism of earthquakes and nature of faulting on the mid-ocean ridges. *J. Geophys. Res.*, **72**, 2131–2153.

Talwani, M. and Ewing, M. 1960. Rapid computation of gravitational attraction of three-dimensional bodies of arbitrary shape. *Geophysics*, **25**, 203–225.

Talwani, M., Le Pichon, X., and Ewing, M. 1965. Crustal structure of the mid-ocean ridges. 2: Computed model from gravity and seismic refraction data. *J. Geophys. Res.*, **70**, 341–352.

Talwani, M., Worzel, J. L., and Landisman, M. 1959. Rapid gravity computations for two-dimensional bodies with application to the Mendocino submarine fracture zone. *J. Geophys. Res.*, **64**, 49–59.

Tapley, B. D., Schutz, B. E., and Eanes, R. J. 1985. Station coordinates, baselines and earth rotation from LAGEOS laser ranging: 1976–1984. *J. Geophys. Res.*, **90**, 9235–9248.

Telford, W. M., Geldart, L. P., and Sheriff, R. E. 1990. *Applied Geophysics*, Cambridge: Cambridge University Press.

Thébault, E., Finlay, C. C. and members of IAGA Working Group V-MOD 2015. International Geomagnetic Reference Field: the 12th generation. *Earth, Planets and Space*, **67**, 79.

Toksöz, M. N., Minear, J. W., and Julian, B. R. 1971. Temperature field and geophysical effects of a downgoing slab. *J. Geophys. Res.*, **76**, 1113–1138.

Torsvik, T. H., Van der Voo, R., Preeden, U., et al. 2012. Phanerozoic polar wander, paleogeography and dynamics. *Earth Sci. Rev.*, **114**, 325–368.

Turcotte, D. L., McAdoo, D. C., and Caldwell, J. G. 1978. An elastic–perfectly plastic analysis of the bending of the lithosphere at a trench. *Tectonophysics*, **47**, 193–205.

Turcotte, D. L. and Schubert, G. 1982. *Geodynamics: Applications of Continuum Physics to Geological Problems*, New York: Wiley.

Turner, G., Enright, M. C., and Cadogan, P. H. 1978. The early history of chondrite parent bodies inferred from ^{40}Ar–^{39}Ar ages. In *Proceedings of the Ninth Lunar and Planetary Science Conference*, Houston, Texas, pp. 989–1025.

Van Andel, T. H. 1992. Seafloor spreading and plate tectonics. In *Understanding the Earth*, Brown, C. J., Hawkesworth, C. J., and Wilson, R. C. L., ed., Cambridge.: Cambridge University Press, pp. 167–186.

Van der Voo, R. 1990. Phanerozoic paleomagnetic poles from Europe and North America and comparisons with continental reconstructions. *Rev. Geophys.*, **28**, 167–206.

Van Nostrand, R. G. and Cook, K. L. 1966. Interpretation of resistivity data. Prof. Paper 499, United States Geological Survey.

Velicogna, I. 2009. Increasing rates of ice mass loss from the Greenland and Antarctic ice sheets revealed by GRACE. *Geophys. Res. Lett.*, **36**, L19503.

Vine, F. J. 1966. Spreading of the ocean floor: new evidence. *Science*, **154**, 1405–1415.

Vine, F. J. and Matthews, D. H. 1963. Magnetic anomalies over oceanic ridges. *Nature*, **199**, 947–949.

Vitorello, I. and Pollack, H. N. 1980. On the variation of continental heat flow with age and the thermal evolution of continents. *J. Geophys. Res.*, **85**, 983–995.

Ward, S. H. 1990. Resistivity and induced polarization methods. In *Geotechnical and Environmental Geophysics, Vol.1*, Ward, S. H., ed., Tulsa, OK: Society of Exploration Geophysicists, pp. 147–190.

Ward, S. N. 1989. Tsunamis. In *Encyclopedia of Solid Earth Geophysics*, James, D. E., ed., New York: Van Nostrand Reinhold, pp. 1279–1292.

Watts, A. B., Cochran, J. R., and Selzer, G. 1975. Gravity anomalies and flexure of the lithosphere: a three-dimensional study of the Great Meteor seamount, Northeast Atlantic. *J. Geophys. Res.*, **80**, 1391–1398.

Weber, R. C., Lin, P.-Y., Garnero, E. J., Williams, Q., and Lognonné, P. 2011. Seismic detection of the lunar core. *Science*, **331**, 309–312.

Wegener, A. 1922. *Die Entstehung der Kontinente und Ozeane* (3rd ed.), Braunschweig: Vieweg.

Wiechert, E. 1897. Ueber die Massenverteilung im Innern der Erde. *Nachr. Ges. Wiss. Göttingen*, 221–243.

Wilde, S. A., Valley, J. W., Peck, W. H., and Graham, C. M. 2001. Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Gyr ago. *Nature*, **409**, 175–178.

Wilson, J. T. 1963. A possible origin of the Hawaiian Islands. *Can. J. Phys.*, **41**, 863–870.

Wilson, J. T. 1965. A new class of faults and their bearing on continental drift. *Nature*, **207**, 907–910.

York, D. and Farquhar, R. M. 1972. *The Earth's Age and Geochronology*, Oxford: Pergamon Press.

Yukutake, T. and Tachinaka, H. 1968. The non-dipole part of the Earth's magnetic field. *Bull. Earthquake Res. Inst. Tokyo*, **46**, 1027–1062.

Zijderveld, J. D. A. 1967. A.C. demagnetization of rocks: analysis of results. In *Methods in Palaeomagnetism*, Collinson, D. W., Creer, K. M., and Runcorn, S. K., ed., Amsterdam: Elsevier, pp. 254–286.

Index

acceleration
  centrifugal, 14, 49, 55, 58
  centripetal, 54, 55, 64
  Coriolis, 57
  Eötvös, 57
  gravitational, 52
  gravity, 88
  tidal, 59
accelerometer, 91, 92
accretion
  Earth and planets, 5, 250
  protoplanets, 15
activation energy, 133, 314
adiabatic condition, 254
adiabatic temperature gradient, 255
advection, 270, 271
age of the Earth
  historical estimates, 237
    biblical, 237
    cooling of the Earth, 238
    cooling of the Sun, 238
    Earth–Moon separation, 238
    oceanic salinity, 239
    sedimentary accumulation, 239
  radiometric
    Precambrian rock ages, 250
    zircon grain ages, 250
age of the Moon, 250
age of the solar system, 250

- alidade, 6
- alkali-vapor magnetometer, 343, 344, 345
- ambient seismic noise, 168, 177
  - continuous random sources, 176
  - Earth's hum, 177
  - interferometry, 178
  - oceanic origin, 178
  - spatial coherence, 177
  - virtual source, 180
- ampère, 289
- anelasticity, 135
  - creep component, 135
  - damping of free oscillations, 158
  - deformation, 134
  - retardation time, 134
  - tidal response, 63
- angular momentum, 2, 3, 4, 12, 65, 238
  - conservation, 3, 7, 57
  - Earth–Moon rotation, 12, 64
  - effect of earthquakes, 64
  - Jupiter, 3
  - of Earth's atmosphere, 64
  - of Earth's core, 64
  - of the planets, 3
- anisotropy
  - magnetic, 364
    - magnetization energy, 360, 361
    - magnetocrystalline, 359, 361, 403
    - magnetostatic, 359, 402
      - demagnetizing factor, 402
    - magnetostrictive, 403
  - seismic, 132, 134, 158
    - elastic properties, 130
    - lattice preferred orientation, 158
    - Love–Rayleigh discrepancy, 159

        - shear-wave splitting, 158
- aphelion, 7
- apogee, 7
- apparent polar wander, 24, 368, 378
- Argand diagram, 106
- asteroids, 1, 5, 14, 15, 16, 18, 250
    - Apollo-type, 15
    - asteroid belt, 6
    - Ceres, 1, 6, 15, 19
    - collisions with Earth, 15
    - collisions with Moon, 250
    - compositional types, 15
    - Earth-crossing orbits, 15
    - Pallas and Vesta, 15
- asthenosphere, 25, 33, 139, 226
- astrolabe, 6
- astronomical unit, 6, 7, 12

- back-arc basin, 33
- barycenter, 18, 58
- basalt
    - ages in island chains, 40
    - classification
        - andesitic type at subduction zones, 33
        - mid-ocean ridge type (MORB), 31
        - oceanic island type, 40, 284
    - magnetic mineralogy, 387
    - oceanic crustal Layer 2A, 30, 386
    - sea-floor magnetization, 29
- Biot–Savart law, 321
- blocking temperature, 364
- Bode’s law, 6, 7
- Bouguer gravity anomaly, 100
    - across crustal block, 100
    - across mountain range, 101
    - oceanic and continental, 113

bow shock, 323, 336
  Earth, 322
  Jupiter, 338
  Mars, 337
  Mercury, 336
  Neptune, 339
  Pluto, 340
  Saturn, 339
  Uranus, 339
  Venus, 337
bulk modulus, 128, 131, 150

Callisto, moon of Jupiter, 8
Chandler wobble, 65, 83, 143
  Love numbers, 62
charge, electrical, 288, 289
  atmospheric, 292
  definition of coulomb, 289
  electron, 289
  field lines, 289
Charon, moon of Pluto, 18
Clairaut's theorem, 75
Clapeyron slope, 231
Clausius–Clapeyron equation, 256
collision hypothesis, origin of planets, 5
complex number, 106
complex plane, 106
concordia–discordia diagram, 248
conduction, electrical, 291
  dielectric, insulator, 291
  electrolytic, 291
  electronic, metal, 291
  mechanisms in the Earth, 314
  semiconductor, crystal, 291
conduction, thermal, 257, 262
  by electrons, 257

  - by phonons, 257
  - crust and lithosphere, 262
  - radiation influence, 260
  - temperature gradient, 281
- conductivity
  - electrical, 308, 311
    - core, 315
    - crust, 314
    - mantle, 312, 314
  - hydraulic, 271
  - thermal, 257, 272
- conjunction of Earth and Moon, 61
- conservation of angular momentum, 3, 7, 57
- conservation of energy, 2, 7, 8
- continental drift, 22, 28, 31
  - paleomagnetic evidence, 24, 369, 380
  - plate tectonics, 24, 380
  - Wegener's hypothesis, 22
- convection, 257, 258, 280
  - compositional, 259
  - mantle models, 283
- core–mantle boundary, 283, 314
  - D″-layer, 226
  - density change, 143
  - diffraction of P-waves, 221
  - electrical conductivity, 314
  - mantle plumes, 40
  - seismic velocity discontinuity, 217, 226
  - temperature, 256
- core–mantle coupling, 64
- Coriolis acceleration, 57
- correlation of time series, 179
- cosmic rays, 241, 248, 292
- Coulomb's law, 239, 288, 289
- creep

  Coble type, 134
  diffusion of atoms, 134
  dislocation mobilization, 133
  Nabarro–Herring type, 134
  role of grain size, 134
crust
  Bouguer gravity anomalies, 113
  Conrad discontinuity, 218
  continental structure, 218
  Mohorovičić discontinuity, 24, 191
  oceanic structure, 114, 115, 352
  thickness, 113, 120
crystal structure
  atomic lattice, 133
  defects and dislocations, 133
Curie temperature, 29, 400
  hematite, 359
  magnetite, 358
  measurement method, 360
  pyrrhotite, 359
  titanohematite, 359
  titanomagnetite, 358
cyclostratigraphy, 68, 391

D'Alembert's principle, 151
D″-layer
  heat flux from core, 284
  influence on seismic waves, 221
  post-perovskite, 285
  screening of secular variation, 315
  superplumes, 285
  thermal boundary, 226, 257
  thickness, 285
Darcy's law, 271
declination, 318, 322
deformation type

  - anelastic, 134
  - brittle, 136
  - ductile, 136
  - elastic, 125
  - plastic, 136
  - shear, 127
  - transitional, brittle–ductile, 136
  - viscoelastic, 135
- density
  - Adams–Williamson equation, 227
  - measurement methods
    - Archimedes’ principle, 97
    - gamma–gamma logging, 98
    - Nettleton profile, 99
    - relationship to seismic velocity, 98, 112
  - variation in the Earth, 227
- dielectric constant, 291, 305, 313
- diffraction, 161
  - Fraunhofer and Fresnel types, 221
- diffusion, 134, 281
  - creep, 134
  - equation, 263, 306
  - ionic, 293
- diffusivity, thermal, 258, 263, 265, 267, 273, 280, 281
- dilatation, 127, 151
- dipole
  - electrical, 291, 305
    - current geometry between electrodes, 299
  - magnetic, 318, 319, 320
    - field geometry, 326, 328
    - potential, 320
- direction cosines, 41, 43
- dislocations, 133
  - creep mechanism, 133
  - glide plane, 133

dispersion
  normal, 155
  surface waves, 154
    phase and group velocities, 155
distributed acoustic sensing (DAS), 187
diurnal variation, 314, 324, 325, 345, 346
Doppler effect, 9, 10, 78, 82
Doppler radar, 10
downward continuation
  geomagnetic field, 327
  gravity anomalies, 106
  magnetic field of Jupiter, 338
dwarf planet, 1, 15, 18, 19
dynamic topography, 75, 76
dynamical ellipticity, 66, 70, 76
dynamical form-factor, J_2, 72
dynamo, geomagnetic, 333

Earth
  age, 237
  density–depth profile, 227
  electrical properties, 292
  gravitational potential, 70
  mass and mean density, 53, 118, 119
  rotation, changes with time, 63
  shape, 49, 68, 70
    flattening, 69, 73, 74, 76
  size, 48
Earth gravitational model 2008, 74
Earth's internal structure, 24, 191, 217
  core
    composition, 226
    diffraction of P-waves at core–mantle boundary, 221
    fluid outer core, 25, 217, 226
    shadow zone, 217, 218, 220
    solid inner core, 226

  - crust
    - Conrad discontinuity, 218
    - continental structure, 229
    - Mohorovičić discontinuity, 217
    - oceanic structure, 228
  - mantle
    - lower, 232
      - D″-layer, 221
      - large low-velocity provinces, 232
    - upper, 229
      - seismic discontinuities, 231
      - temperature anomalies, 230
      - velocity anomalies, 230
  - seismic phases, 219
    - ray paths, 218
    - travel times, 219
  - velocity models, 225
    - *iasp 91*, 225, 226
    - PREM, 225
- Earth’s orbit
  - dimensions, 11
  - eccentricity, 12
    - cyclical change, 67
  - precession of perihelion, 67
- Earth–Moon rotation
  - barycenter, 58
  - future synchroneity, 65
  - tidal lag, 63
- earthquake characteristics
  - aftershocks, 193
  - duration of ground motion, 206
  - elastic rebound model, 191, 192, 205
  - energy and magnitude, 208
  - epicenter location, 193, 194
    - Wadati diagram, 194

  - epicentral distance, 193
  - focal depth, 193, 194
  - focal mechanism, 33, 196
    - double-couple source, 197
    - first motion quadrants, 197
    - radiation pattern, 197
    - single-couple source, 197
  - focus (hypocenter), 193, 196
  - foreshocks, 210
  - frequency of occurrence, 208
  - global distribution of epicenters, 25, 194
  - ruptured area, 206
  - secondary effects, 213
    - conflagrations, 215
    - landslides, 213
    - tsunami, 213
  - slow-slip events, 192
  - swarm events, 192
- earthquake intensity, 203
  - European Macroseismic Scale 1998, 204
  - isoseismal maps, 204
  - modified Mercalli scale, 203
  - relationship to magnitude, 203, 207
  - Rossi–Forel scale, 203
- earthquake magnitude, 205
  - body-wave magnitude, 205
  - moment magnitude, 206
  - Richter scale, 205
  - saturation effect, 205
  - seismic moment, 206, 207
    - centroid moment tensor analysis, 207
  - surface-wave magnitude, 205, 206
- earthquakes and society
  - early warning, 213
  - hazard analysis, 210

    - ground acceleration, 210
    - ground motion models, 210
    - probabilistic model, 211
    - seismic hazard maps, 211
  - nuclear explosion monitoring, 216
    - discrimination criteria, 217
  - prediction
    - Haicheng, China, success, 209
    - Parkfield experiment, 209
  - prediction difficulties, 209, 210
- earthquakes, historical, 206
  - Agadir (1960), 208
  - Aleutian (1946), 214
  - Anchorage (1964), 205
  - Chile (1964), 215
  - Chimbote, Peru (1970), 213
  - Loma Prieta (1989), 205
  - Mexico City (1985), 205
  - New Madrid (1811–1812), 195
  - San Francisco (1906), 205
  - Sumatra–Andaman (1960), 207, 213
  - Tangshan (1976), 193
  - Tohoku (2011), 213, 215
- East African Rift Zone, 35
- ecliptic plane, 1, 3, 8
- eclogite, 279
- eikonal equation, 152, 159
- elasticity, 125
  - elastic limit, 135
  - elastic moduli, 128
  - Hooke’s law, 125
- electrical surveying methods, 292
  - electromagnetic induction, 307
    - airborne surveying, 308
    - depth penetration, 308

        - horizontal loop (HLEM), 308
        - in-phase and quadrature responses, 308
        - multi-frequency technique, 308
    - ground penetrating radar, 312
    - induced polarization, 302
        - decay time, 303
        - field method, 302
        - ionic accumulation, 302
        - pseudo-section, 303
    - magnetotelluric, 310
    - resistivity, 296
        - double-dipole configuration, 298
        - Schlumberger configuration, 298
        - Wenner configuration, 298
    - resistivity tomography, 304
    - self-potential, 293
        - anomaly, 294
        - galvanic action, 294
        - oxidation (redox) potential, 294
        - survey methods, 294
            - potential gradient, 294
            - total field, 295
    - telluric currents, 295, 296
    - transient electromagnetic, 309
        - airborne surveys, 310
        - secondary field decay curve, 309
        - skin depth effect, 309
    - vertical sounding, 300
- electricity
    - electric charge, 289
    - fundamental principles, 289
    - historical concepts, 288
    - measurement units, 290
- electromagnetism, 304
    - electromagnetic wave, 305

  - frequency spectrum, 305
  - induction surveying, 305, 306
  - induction theory, 306, 307
  - Maxwell's equations, 304, 305
- Eötvös acceleration, 57, 93
- epicenter, 193
- equinoxes, 66
  - definition, 8
  - precession of, 66, 67, 236
- equipotential surface, 54, 59, 62, 73, 75, 80
- error function, 265, 266
- Euler pole of rotation, 28, 34, 35, 41, 42
  - fit of Atlantic coastlines, 42
  - method of location, 42
  - paleogeographic reconstruction, 378, 380, 381
  - plate motion on sphere, 42
- Euler, Leonhard
  - complex numbers, 106
  - rotation on a sphere, 41
  - theory of free nutation, 65
- Europa, moon of Jupiter, 16
- European Geotraverse (EGT), 114
- European Radar Satellites (ERS), 82
- exoplanets, 7

- Farallon plate, 37
- fault-plane solutions, 197
  - beachball plot, 198
  - continental collision zones, 199, 201
    - Alps, 203
    - Himalaya, 201
  - fields of compression and dilatation, 198
  - oblique focal mechanism, 201
  - oceanic ridge, 199, 200
  - stereogram representation, 197
  - subduction zone, 201, 202

- intermediate and deep earthquakes, 201
- shallow earthquakes, 201
- transform fault, 200

faults
- mechanics of faulting, 198
- plane of failure, 198
- stress axes, 198
- tectonic faults, 199
  - listric faults, 199
  - normal faults, 199
  - reverse (thrust) faults, 199
  - transcurrent (strike–slip) faults, 80, 192, 199
  - transform faults, 199

Fermat’s principle, 161
- reflection law, 162
- refraction law, 162

ferrimagnetism, 357, 401

ferromagnetism, 357, 400

field
- concept, 289
- electric, 289, 290
- electromagnetic, 304
- gravitational, 52
- magnetic, 318, 320

figure of the Earth, 68, 69, 73, 80, 112

finite-difference analysis, 173
- approximation of derivatives, 174
- discretization, 173
- discretized wave equation, 174
- iterative solution, 174

flexural parameter, 138, 139

flexural rigidity, 137, 138, 139, 141

flexure of lithosphere
- at oceanic islands, 137
- at subduction zones, 78, 116, 138

  - Great Meteor seamount, 138
- flow
  - laminar, 132
  - plastic, 135
  - solid-state, 132, 133
  - viscous, 132
- Fourier
  - analysis, 104, 309, 371
  - integral, 105
  - series, 104
  - series, double, 105
  - theory of heat conduction, 262
  - transform, 107
- fracture zones, 26, 34
- free-air gravity correction, 97

- Gal, definition, 52
- gamma–gamma density logging, 98
- Ganymede, moon of Jupiter, 16, 339
- Gauss coefficients, 327
- generation of internal field
  - dynamo theory
    - core conditions, 333
    - frozen-flux theorem, 333
    - magnetohydrodynamic equation, 333
    - simulations, reversing field, 334
  - historical models, 332
  - Precambrian geodynamo, 333
- geoid
  - definition, 75
  - residual, correlation with hotspots, 38
  - secular changes, 81
  - undulations, 75, 76, 78
    - large-scale, 76
- geomagnetism
  - core surface field, 327, 328, 331

  - dip poles, 329
  - dipole field, 328
  - dipole magnetic moment, 329
  - field elements, 319, 322, 325
  - field of internal origin
    - potential, 326
    - spherical harmonic analysis, 324
  - geomagnetic poles, 329
  - impulses (jerks), 314, 332
  - magnetopause, 323
  - magnetosheath, 323
  - magnetosphere, 322
  - magnetotail, 323
  - multipole representation, 326
  - non-dipole field, 330
  - potential of total field, 322
  - power spectrum, 327, 332
  - satellite measurements
    - CHAMP mission, 325
    - MAGSAT mission, 325
    - NGDC-720 model field, 327
    - SWARM mission, 325, 327
  - secular variation, 331, *See also* secular variation
  - separation of external and internal fields, 322
  - time spectrum of variations, 314
- geophone, 184
- geopotential, 73, 75
- georadar, 313
- geothermal gradient, 258
- Giant Impact hypothesis of lunar origin, 335
- Global Navigation Satellite System (GNSS), 66, 79
- global positioning system (GPS), 79
  - accuracy of location, 79
  - GPS networks, 79
  - plate tectonic motions, 79

Gondwana, 23, 24, 380
  break-up, 383
  early models, 22
  reconstruction, 22, 23
gradiometer
  gravity, 80, 87, 91
  magnetic, 325, 345
GRAIL lunar mission, 13
gravimeter, 89
  astatized, 89
  instrumental drift, 95
  spring type, 90
  superconducting type, 90
  zero-length spring, 89
gravitation
  geocentric constant, *GE*, 52
  gravitational constant, *G*, 9, 51
    best estimate, 51
    definition, 51
    measurements, 51
  gravitational field, 52
  gravitational potential, 52, 70
  inverse square law, 56
gravity
  corrections to measurements, 93
    Bouguer plate, 94, 96
    free-air, 94
    latitude, 95
    time-dependent, 95
    topographic (terrain), 94, 95, 96
  gradient matrix and tensor, 93
  gradiometry, 91
  gravimetry. *See* gravimeter
  increase from equator to pole, 74
  measurement methods

    - absolute, 87
        - free-fall method, 88
        - pendulum method, 87
        - rise-and-fall method, 88
        - worldwide gravity network, 89
    - borehole method, 99
    - marine and airborne methods, 91, 93
    - relative measurement, 89
  - normal gravity, 73
  - radial variation in the Earth, 227
  - survey methods, 91
  - units and subunits, 52
  - values on reference ellipsoid
    - normal gravity formula, 74
    - Somigliana formula, 74
- gravity anomalies, 32
  - analysis
    - downward continuation, 105
    - Fourier series representation, 103
    - polynomial representation, 102
    - regional and residual anomalies, 102
    - visual analysis, 102
    - wavelength filtering, 105
  - Bouguer anomaly, 100
  - crustal structure, 112, 113
  - density contrast, 101
  - free-air anomaly, 100
  - isostatic anomaly, 122
  - modeling
    - anticline and syncline, 109
    - general 3D body, 112
    - geometric models, 106
      - horizontal cylinder (anticline/syncline), 109
      - horizontal line element, 108
      - horizontal slab (vertical fault), 111

      - horizontal thin sheet, 110
      - sphere (diapir), 107
    - iterative modeling, 111
  - mountain ranges, 113, 114
  - oceanic ridges, 114
  - subduction zones, 116
  - wavelength, 101
- ground penetrating radar, 312
- group velocity, 155, 214
- Grüneisen parameter, 256
- gyromagnetic ratio, 344

- heat, 253, 254
  - conduction equation, 264
  - depth penetration of external heat, 265
  - mechanical equivalent, 254
  - sources in the Earth, 260
- heat flow, 254
  - continental, 261, 267
    - borehole measurements, 268
    - effect of porosity, 270
    - variation with crustal age, 269
  - equation, 257, 262
  - global variation, 273, 274, 275
  - mean heat-flow values by age, 274
  - plate tectonics
    - back-arc basins, 33, 260, 279
    - cooling of lithosphere, 265, 267
    - deep-sea trench, 33
    - oceanic, 271, 275
      - effect of hydrothermal circulation, 273
      - ridge systems, 32, 260
      - variation with lithospheric age, 272
    - subduction zone, 33, 279
  - radiogenic production, 261
    - heat-flow provinces, 261

    - reduced heat flow, 261
  - sources in crust and mantle, 262
  - spherical harmonic analysis, 275
- Hooke's law, 89, 125, 128, 131, 135
  - elastic deformation, 125
  - elastic moduli, 128
  - linearity and elastic limits, 135
- hotspots, 38, 40, 286
  - elevation of geoid, 38
  - global distribution, 38
  - reference frame, 40
  - type of basalt, 40
- Huygens' principle, 159, 187
  - seismic diffraction, 161
  - seismic reflection, 160
  - seismic refraction, 160
- hydrostatic pressure, 228

- impact hypothesis
  - Cretaceous–Tertiary extinctions, 15
  - lunar origin, 12
  - power for early geodynamo, 383
- inclination, magnetic, 322
  - archeomagnetic field, 369
  - effect of true polar wander, 381
  - field direction, 318, 328, 329
  - flattening error, 365, 366, 370
  - GAD hypothesis test, 370
  - polarity reversal, 384
  - rock magnetization direction, 353, 376, 377
  - sign convention, 371
- inclination, orbital, 8
  - Earth satellites, 77, 80
  - Pluto's anomalous orbit, 18
  - the Moon, 12, 67
- incompressibility, 128

inertia, 54
InSAR technique, 82
  interference fringes, 83
insolation, 67
International Geomagnetic Reference Field, 325, 340, 344
International Reference Ellipsoid, 69, 73
inversion of seismic data
  direct method, 221
    Herglotz–Wiechert inversion, 222
    travel times, 222
  indirect method, 221
    forward problem, 221
    seismic tomography, 222
ionosphere, 5, 293, 295, 310, 311, 314, 323, 346
  electrical currents, 324
  Kennelly–Heaviside layer, 324
  layered structure, 324
  magnetic storm source, 324
island arcs, 26, 194
  gravity anomalies, 116, 120
  isostatic imbalance, 113
  locations at subduction zones, 31, 33
  marginal basins, 260, 280
isochron
  diagram, 246, 248
  mineral versus whole-rock ages, 244
  radiometric, 244, 246
  sea-floor spreading lineations, 390
isostasy, 118, 119
  hydrostatic equilibrium, 121, 122
  isostatic adjustment, 81, 137
    Fennoscandia, 121
    northern Canada, 121
    Swiss Alps, 123
  models

    - Airy, buoyancy, 119
    - Pratt, buoyancy, 120
    - Vening Meinesz, elastic plate, 120, 122
  - mountain range root-zones, 119
  - oceanic lithosphere equilibrium, 278
- isotherms
  - cooling lithospheric plate, 276, 277, 279
  - subducting cold plate, 279, 280
  - topographic influence on heat flow, 268
- isotopes, 239, 260
  - beryllium, 248
  - cosmogenic, 248
  - oxygen, 241
  - primordial, 241
  - radioactive krypton, 249
  - stable, 241
- isotopic ratios, 244
  - ion microprobe, 243
  - isochron diagrams, 244, 246, 248
  - mass spectrometer, 242
  - oceanic island basalts, 284
  - oxygen isotopes, 249
  - radiometric dating, 245
  - use for cyclostratigraphy, 386

- Jupiter, 15
  - atmosphere, 16
  - Europa, 16
  - Galilean moons, 16, 339
  - gravitational field, 15
  - internal structure, 16
  - Juno mission, 15, 338
  - magnetic field, 16, 338
  - mass and angular momentum, 3
  - ring system, 16

Kepler’s laws of planetary motion, 7, 9, 12
  inverse square law of gravitation, 56
Königsberger ratio, 340
  continental crustal rocks, 341
  oceanic basalts, 340, 353
Kuiper belt, 18

Lamé constants, 130
  relationships to elastic constants, 131
Laplace equation, 71, 72
large low-velocity province (LLVP), 232
Larmor precession, 342, 344, 399
laser ablation, 243
Laser Geodynamics Satellites (LAGEOS), 77
laser-ranging, 9, 12, 52, 66, 77
Laurasia, 24, 383
Laurussia, 22, 23, 378, 382, 383
Legendre equation, 71
Legendre polynomials, 71, 72, 75, 325
Lehmann discontinuity, 225
length of the day, 83
  Babylonian and Arab datasets, 63
  effect of the atmosphere, 64
  effect of tidal friction, 63
lightning, 292
lineations, magnetic, 28, 29, 30, 34, 387
lithosphere, 25
  continental, 31
  flexure, 116, 137, 138
    linear load, 138
    thin plate model, 138
    three-dimensional load, 138
  oceanic
    cooling models, 275
      boundary-layer model, 277
      flat plate model, 277
      half-space model, 276
      other heat sources, 277
    thermal structure, 277
    thickness variation with age, 31, 277
  rigidity, 137
  tectonic plates, 25

        - size distribution, 284
    - thickness, 140
- Lorentz field, 243, 321, 333, 335
- Love numbers, 62
    - Chandler wobble, 62
    - Mercury, 63
    - tidal theory, 62
- Love wave, 154
    - particle motion, 154
    - velocity, 154
- Lunar Laser Ranging experiment, 9, 12

- MacCullagh's formula, 70
- magnetic anomalies, 346
    - geometric models, 349
        - anticline/syncline, 350
        - horizontal crustal block, 350
            - anomaly shape, 351, 353
        - Poisson's relation, 351
        - sphere, 349
        - vertical dike, 348
            - approximation by dipole, 349
            - effect of depth extent, 348
    - inclined magnetization, 349
        - reduction to the pole, 349, 353
        - skewness of anomaly shape, 349
    - oceanic lineations, 37, 352
        - anomaly shapes, 353
        - multiple source layers, 353
    - surface distribution of poles, 346, 348
    - wavelength of anomaly, 327
- magnetic fields, extraterrestrial, 336
    - Jupiter, 338
    - Mars, 337
    - Mercury, 336
    - Neptune, 339

- Pluto, 339
- Saturn, 339
- the Moon, 335
  - measurement using Lorentz force, 335
- the Sun, 334, 335
- Uranus, 339
- Venus, 337

magnetic minerals
- ferrimagnetic mineral identification, 359
- goethite, 359
- greigite, 359
- hematite, 359
- maghemite, 359
- magnetic properties
  - grain size dependence, 360
    - first-order reversal analysis, 363
    - multidomain behavior, 362
    - single-domain behavior, 361
    - stability of magnetization, 361
- magnetite, 358
  - biogenic formation, 368
- pyrrhotite, 359
- ternary oxides, 358
- titanohematite series, 359
- titanomagnetite series, 358

magnetic relaxation, 361

magnetic storms, 314, 324, 325, 335

magnetic surveying, 340
- airborne method, 344
- Königsberger ratio, 340
- marine method, 345
- reduction of measurements, 346
  - altitude correction, 346
  - latitude correction, 346
- reference field, 340, 346

  - survey design, 345
- magnetic susceptibility, 357, 358
  - Chinese loess profiles, 367
  - paleoclimatic indicator, 368
- magnetism
  - biogenic, 368
  - discovery, 317
  - environmental, 367
  - flux density, 317
  - magnetic B-field, 320
  - magnetic H-field, 357
  - magnetic poles, 318, 319, 320
  - physics of, 321
  - pioneering discoveries, 317, 318
- magnetization, 356
  - Earth’s crust, 340
  - induced, 341
  - remanent, 340
    - analysis of components, 373
      - vector diagram, 373
    - anysteretic (ARM), 367
    - characteristic (ChRM), 372
    - chemical (CRM), 366
    - depositional (DRM), 365
    - isothermal (IRM), 366, 400
    - natural (NRM), 363
    - thermal (TRM), 364
    - viscous (VRM), 366
  - spontaneous, 400
- magnetometer
  - absolute measurement, 342
  - absorption-cell (optically pumped), 343
  - cryogenic, 371
  - flux-gate, 336, 341
  - optical pumping, 343

  - proton-precession, 342
    - gyromagnetic ratio, 342
  - scalar type, 342
  - vector type, 336, 342
- magnetopause, 323, 336
- magnetosheath, 323
- magnetosphere, 322
- magnetostratigraphy, 385
  - dating by cyclostratigraphy, 386
  - deep-sea cores, 385
  - polarity timescales, 385
  - radiometrically dated lavas, 385
- magnetostriction, 403
- magnetotactic bacteria, 368
- magnetotail, 323
- magnetotelluric sounding, 310
  - deep conductivity structure, 311
  - electric and magnetic components, 310
  - interpretation, 311
  - inversion of data, 312
- mantle
  - convection, 280
    - cell aspect ratio, 283, 284
    - D″-layer, 285
    - layered mantle model, 283, 284
    - physical parameters, 282
    - sizes of lithospheric plates, 284
    - superplumes, 285
    - thermally activated creep, 284
    - whole mantle model, 284
  - plumes, 230, 285
    - definition of feature, 286
    - difficulties with concept, 286
    - relationship to D″-layer, 226
    - relationship to hotspots, 40

    - role in mantle convection, 40
  - structure
    - D″-layer, 226
    - large low-velocity province (LLVP), 232
    - phase transitions, 225, 226, 257
    - upper mantle transition zone, 225
    - velocity discontinuities, 218, 226, 231
  - viscosity, 140, 140, 141, 143, 259
    - effects on Earth rotation, 142
    - estimates, 142, 142
- Mars, 14
  - crustal magnetic anomalies, 338
  - exploration with satellites and landers, 14
  - impact craters, 14, 338
  - internal structure, 14
  - magnetic field, 337
- mascon, 13
- mass spectrometer, 242
  - inductively coupled plasma, 243
  - ion microprobe, 243
  - laser ablation, 243
- melting point gradient, 256
- Mercury, 10, 336
  - gravity field, 63
  - internal structure, 11, 336
  - magnetic field, 336
  - MESSENGER mission, 11, 63, 336
  - periods of spin and orbit, 11
- meteorites, 249, 250
  - Allende, age of solar system, 250
  - compositional types, 250
- microseisms, 176
- Milankovitch climatic cycles, 67
  - periodicities, 68
- mineral phase transitions, 257

- lower mantle, 285
- upper mantle, 231, 255, 279

Mohorovičić discontinuity, 24, 191, 217, 218
- seismological and petrological, 218

moment of inertia
- axis of figure, 65
- dynamical ellipticity, 70
- MacCullagh's formula, 70
- planetary, 2, 4
- point mass, 2
- solid body, 2
- spherical shell, 20, 85
- uniform sphere, 86

Moon, 12
- distance from the Earth, 12
- early history, 335
- Giant Impact hypothesis, 12, 250, 335
- GRAIL mission, 14
- gravitational field, 13
- internal structure, 14, 335
- libration of longitude, 12
- moonquakes, 14
- revolution without rotation, 58
- seismic velocity-depth profiles, 14

nebular hypothesis, 3, 5

Néel temperature, 401

Neptune, 17, 339
- internal structure, 17
- magnetic field, 339
- rings and moons, 17

Newton, Isaac, 50
- calculation of relative mass of Jupiter, 8
- curved path of projectile, 54
- law of universal gravitation, 50
- laws of motion, 50

  spheroidal shape of Earth, 49, 50
  verification of law of gravitation, 56
normal distribution (bell curve), 266
Nusselt number, 282
nutation, 15
  forced, 65, 66
  free, 65

obliquity of rotation axis, 8, 10, 67, 236, 339
  cyclical changes, 67
  Earth, 12
  Neptune, 17
  Saturn, 16
  Uranus, 17
occultations, defining planet's size, 9, 11, 14, 16, 63
ocean depth increase with lithospheric age, 272, 273, 275, 277, 278
octupole field, 326
Ohm's law, 290
Oort Cloud, 19
opposition of Earth and Moon, 61
oscillations, free (natural), 155
  attenuation with depth, 157
  comparison with surface waves, 157
  normal modes, 91, 156
  overtones, 156
  power spectrum, 158
  radial oscillations, 156
  spherical oscillations, 156
  toroidal oscillations, 157
oxygen isotope ratio, 241
  ice cores, 249
  paleotemperature proxy, 241
  temperature dependence, 241

paleoclimate
  evidence for continental drift, 22, 368, 380

  - magnetic proxies, 367
  - Milankovitch cycles, 67
  - oxygen isotope ratio, 241, 242
- paleomagnetism, 23, 356
  - analysis of magnetization
    - component separation, 371
    - progressive demagnetization, 371
    - stability of components, 372
  - analysis of remanent magnetization
    - progressive demagnetization, 372
    - vector diagram, 373
  - apparent polar wander, 378
    - correlation of APW paths, 378, 379
  - evidence for continental drift, 23, 380
  - field tests of stability, 374
    - conglomerate test, 375
    - fold test, 374
    - reversals test, 375
  - Fisher statistics of directions, 372
    - 95% angle of confidence, 374
    - precision parameter, 373
  - geocentric axial dipole hypothesis, 369, 370, 371, 376
  - paleogeographic reconstructions, 380
    - Paleozoic supercontinents, 381
      - Gondwana, 383
      - Laurussia, 382
      - Pangea, 382
  - paleomagnetic pole, 378
  - Precambrian paleomagnetism, 383
  - rock magnetometers, 371
  - tectonic applications, 375
  - true polar wander, 380, 381
    - hotspot APW, 381
  - virtual geomagnetic pole location, 376, 377
- Pangea, 22

  - break-up, 30, 383, 388
  - duration of supercontinent, 23
  - formation, 380, 382
- parallax method, 6
- perigee, 7
- perihelion, 7, 67
- permeability
  - hydraulic, 216, 270
    - darcy unit, 271
  - magnetic, 305, 320, 357
- permittivity constant, electrical, 289
- phase transitions
  - upper mantle, 230
- phase velocity, 151, 155, 214
- phonons, 257
- planetesimals, 5
- planets
  - angular momentum, 4
  - characteristic properties, 3
    - gaseous giants, 1, 5
    - rocky terrestrial, 1, 5
  - definition of a planet, 1
  - dimensions, 2, 9
  - orbital parameters, 3
  - ring systems, 16
  - theory of origin, 5
- plasma, 322
- plastic flow, 135
- plate tectonics, 25, 31
  - continent–continent collision, 31
    - suture zone, 31, 137
  - driving forces, 43, 44
  - models of plate motions, 30
  - plate margins, 28, 199
    - conservative, transform faults, 34

        - constructive, oceanic ridges, 31
        - destructive, subduction zones, 32
        - diffuse boundaries, 35
        - seismic zones, 28, 195
    - role of triple junctions, 35
    - spreading rates
        - geological observations, 31
        - satellite observations, 77
    - volcanic island chains, 38
- plumb-line, 54, 94
    - anomalous deflections near mountains, 119
    - deflections near mountains, 118, 119
- Pluto, 1, 8, 18
    - anomalous orbit, 18
    - internal structure, 18
    - New Horizons flyby mission, 18
- Poisson solid, 131
- Poisson's ratio, 127, 128, 129, 131
- Poisson's relation, 351
- polar flattening, 2
    - Earth, 69, 73, 74, 76
    - Mars, 14
    - Neptune, 17
    - Saturn, 16
    - Uranus, 17
- polar motion
    - APW paths, 379
    - Chandler wobble, 65
    - effect of isostatic adjustment, 143
    - polhody, 65
    - satellite tracking, 77
- polarity, paleomagnetic
    - chrons and subchrons, 384, 385
    - historical studies, 383, 384
    - magnetostratigraphy, 385

  - oceanic magnetic anomalies, 386
    - Vine–Matthews–Morley hypothesis, 353, 387
  - reversals, 29, 375
  - timescales, 30, 385
  - transitions, 384
- polhody, 65
- porosity, 98, 270
  - electrical conductivity, 293
  - grain size effect, 270
  - influence on induced polarization, 304
- potential
  - centrifugal, 55
  - difference (volt), 290
  - dipole, 319
  - electrical, 289, 297
  - geomagnetic, field of internal origin, 325
  - geomagnetic, total field, 322
  - gravitational, 52, 70
  - logarithmic, 109
  - single magnetic pole, 320
  - theory, Laplace equation, 71
- Prandtl number, 281
- precession
  - Earth's orbit, 67
  - Earth's rotation axis, 66, 67, 76, 236
  - nuclear magnetometers, 327
  - satellite orbit, 77
- Preliminary Reference Earth Model, 225
- pressure, radial variation in the Earth, 228
- protoplanets, 15
- P-wave, 149

- quadrupole field, 326

- radar, 9
  - ranging, 9, 10

  - synthetic aperture, 82
  - wavelengths, 9
- radiation, 257, 259
  - black-body, 260
  - Stefan’s law, 260
- radioactivity, 239
  - bequerel, 240
  - decay process, 240
  - decay rate, 240
  - half-life, 240
  - heat production, 260, 261, 262
  - nuclear activity, 240
  - nuclear structure, 239
- radiometric dating
  - environmental processes, 248
    - beryllium, 248
    - noble gases, 249
  - methods, 242
    - argon–argon, 245
    - lead–lead, 248
    - potassium–argon, 244
    - radiocarbon, 242, 248
    - rubidium–strontium, 244
    - uranium–lead, 246
- Rayleigh number, 258, 281
  - critical value for convection, 259
  - mantle convection, 281, 282
  - radiogenic convection, 281, 282
  - thermal convection, 282
- Rayleigh wave, 154
  - particle motion, 154
  - velocity, 154, 155
- Rayleigh–Bénard convection, 281, 282
  - cell aspect ratio, 283, 284
- receiver function analysis, 221

relaxation time
  isostatic uplift, 140, 141
  magnetization, 360, 361, 364, 366
resistivity, 290, 292
  apparent, 300
  electrolytic, 293
  electronic, 291
  semiconductor, 291
  surveying, 296
    electrode configurations, 298
    lateral profiling, 298
    vertical sounding, 298
  tomography, 304
retroreflector, 12, 13
  absolute gravity measurement, 88, 89
  LAGEOS satellite, 77
  principle, 13
  satellite laser ranging, 77
revolution without rotation, 58, 59
Reynolds number, 282
ridges, oceanic, 26
  heat flow, 32
  sea-floor spreading, 28
  seismicity, 32
  spreading centers, 31
  spreading rates, 30
rigidity modulus, 128, 133
Roche limit, 59, 60, 238
rock magnetism, 356
  magnetic domains, 341
  magnetic minerals, 358
  remanent magnetizations, 363
  types of magnetic behavior, 357
Rodrigues formula, 71

satellite geodesy, 76

  - dynamic oceanic topography, 81
  - gravity and geoid measurement
    - CHAMP mission, 80
    - GOCE mission, 81
    - GRACE mission, 80
    - GRACE-FO mission, 80
  - mass loss in ice-sheets, 81
  - potential coefficients J_n, 72
  - satellite altimetry, 77
  - satellite laser-ranging (SLR), 77
    - determination of *GE*, 52
    - plate tectonic spreading rates, 77
  - time-dependent mass exchanges, 81
- Saturn, 16
  - composition, 17
  - internal structure, 17
  - magnetosphere, 339
  - moons, 16
  - rings, 16
- scattered disk objects, 19
- sea-floor spreading, 28, 29, 265, 353
  - anomaly pattern, 390
  - back-arc basins, 260, 280
  - Central Atlantic, 391
  - half-rates at ocean ridges, 29
    - GPS measurements, 31
    - uniformity of spreading rate, 387
  - initiation of current episode, 388
  - polarity reversal record, 387
- secular variation, 314, 330
  - dipole field, 330
    - archeomagnetism, 331
    - historic intensity decrease, 330
    - location of dipole axis, 331
  - fluid flow at core surface, 332

- non-dipole field, 331
  - archeomagnetic evidence, 331
  - westward drift, 331

seismic arrays
- enhanced detection of events, 216
- nuclear explosion monitoring, 216
- proof of solid inner core, 218
- records of lunar seismicity, 14

seismic tomography
- indirect inversion, 222
- travel-time anomalies, 224
- travel-time residuals, 223
- velocity anomalies, 223, 224
- waveform tomography, 224

seismic velocity discontinuities, 143
- Conrad discontinuity, 218, 229
- core–mantle boundary, 226
- D″-layer, 226
- density changes, 227
- Gutenberg discontinuity, 217, 226
- inner core–outer core, 226
- Lehmann discontinuity, 225
- Mohorovičić discontinuity, 218
- upper mantle transitions, 225, 283

seismic wave propagation, 147
- attenuation, 153
  - geometric spreading, 153
  - viscoelastic damping
    - absorption coefficient, 153
    - quality factor, 153
- body waves, 147
  - energy partitioning at interface, 163, 167
  - P-waves, 148
  - reflection and refraction coefficients, 167, 168
  - seismic impedance, 167

    - S-waves, 149
    - travel times to epicentral distances, 219
    - velocity–depth models, 225
  - energy density of wave, 152
  - spherical wave, 147
  - surface waves, 147, 154
    - dispersion, 154
    - Love waves, 154
    - Rayleigh waves, 154
  - wavefront, 147, 148
- seismicity
  - global
    - active seismic zones, 25, 194
    - focal depths, 195
    - intraplate, 195
    - subduction zones, 33
    - tectonic plate margins, 25
    - Wadati–Benioff zone, 196
  - induced, 215, 216
    - by fluid injection, 215
    - by gas extraction, 216
    - by hydraulic fracturing, 216
- seismology
  - Fermat's principle, 162
  - Huygens' principle, 160
  - ray paths
    - concentric uniform shells, 219
      - epicentral distances, 220
    - curvature of path, 172
    - multi-layered Earth
      - epicentral distances, 220
    - shallow layer reverberations, 221
  - receiver function analysis, 221
  - reflection seismology, 164, 167, 169
    - continental crustal section, 169

      - field procedures, 184
        - common mid-point method, 185
        - continuous profiling, 185
        - split-spread method, 185
      - inclined reflector, 165
        - depth falsification, 166
      - layer thickness estimation, 165
      - migration of section, 167
      - noise sources, 168
        - ground roll, 169
        - multiple reflections, 169
      - normal moveout, 164, 165
      - stacking of records, 169
      - supercritical reflections, 164
      - syncline and anticline, 166
      - two-way travel-time, 164
    - refraction seismology, 169
      - apparent velocities, 172
      - critical angle of incidence, 163
      - critical distance, 164
      - crossover distance, 170
      - double refracted ray, 170
      - head wave, 169
      - inclined interface, 171
      - intercept time, 170
      - travel time, 170
    - seismic noise. *See also* ambient seismic noise
    - seismogram, 147
      - seismic phases, 149
      - synthetic seismogram, 167, 207
- seismometer
  - broadband, 184
    - arrays, 185, 185
  - equation of motion, 182, 183
  - floating type, 187

    - inertial type
        - electromagnetic, 181
        - horizontal motion, 181
        - principle, 180
        - vertical motion, 181
    - instrumental damping, 182
    - invention and development, 191
    - long-period, 184
    - ocean-bottom, 186
    - optical, distributed acoustic sensing, 187
    - resonant frequency, 182
    - short-period, 184
- separation of variables, 263, 264
- shear modulus, 128, 129, 150
- skin depth, 307, 309
- Snell’s law, 161, 163, 164, 170, 172
- solar system
    - geocentric model, 5
    - heliocentric model, 5, 7, 11
    - theories of origin, 3
- solar wind, 322
- solid angle, 347
- solstices, 8
- Somigliana’s normal gravity formula, 74
- spherical harmonic analysis, 318
    - free oscillations, 156
    - geoid potential, 75
    - geomagnetic field, 322, 327, 369
        - Gauss–Schmidt coefficients, 326
    - global heat flow, 275
    - gravitational field, 74
    - Jupiter’s magnetic field, 338
- spherical harmonic functions, 72, 324
    - sectorial and tesseral harmonics, 72, 326
    - zonal harmonics, 326

spherical trigonometry, 42
stacking
  of correlation functions, 178, 180
  of seismic records, 169
standing wave, 156, 157
Stefan–Boltzmann constant, 260
stereogram
  first motions in earthquake, 198
  paleomagnetic directions, 371
strain
  extension, 126
  longitudinal, 126
  matrix, 126, 128
  shear, 127, 128, 150, 395
strain-hardening, 136
stress
  deviatoric, 198
  lithostatic or hydrostatic, 198
  matrix, 126
  normal, 126, 128, 129, 131
  shear, 126, 132, 133, 150, 198
subduction zone, 32, 138
  forces driving plate tectonics, 43
  gravity anomalies, 34, 116
  heat flow, 33
  island arcs, 33
  lithospheric flexure, 137
  role of cold descending slab, 44
  seismicity, 33
  trench suction, 44, 45, 284
  volcanism, 33
Sun, 334
  composition, 334
  convection, 334
  magnetic field, 334

  - photosphere, chromosphere, corona, 334
  - rotation, 334
  - solar flares, 335
  - sunspots, 334
    - magnetic fields, 335
- superconductivity, 90
- superparamagnetism, 361
- superplumes, 232, 285
- susceptibility
  - antiferromagnetic, 401
  - definition, 340
  - diamagnetic, 399
  - paramagnetic, 399, 400
    - Curie–Weiss law, 399
- S-wave, 149
- synthetic aperture radar, 82
  - interferometric SAR, 82

- telluric currents, 295
- temperature
  - absolute zero, 254
  - concept, 253
  - measurement scales, 254
  - profile in the Earth, 255, 256
- temperature gradient, 238
  - adiabatic, 255, 257, 258
  - continental borehole measurements, 267
  - in D″-layer, 284
  - geothermal, 258
  - measurement in boreholes, 267
  - measurement in the oceans, 272
  - melting point, 256
  - outer core, 333
  - subduction zone, 279
  - superadiabatic, 258, 259, 281, 282
- tesla, 319, 321

Theia, 12, 383
thermal boundary layers, 226, 257, 259, 277, 284
thermally activated process
  Arrhenius equation, 361
  creep, 133, 284
  diffusion, 246
  lithosphere ductility, 136
  semi-conduction, 314
thermistor, 267, 272
thermochemical piles, 232
thermodynamic parameters, 253, 254
tides
  anelastic response, 63
  bodily Earth-tides, 57, 62
  diurnal inequality, 59
  effect on gravity measurements, 61
  horizontal tide, 62
  increase in Earth–Moon distance, 64
  increase in length of day, 63
  Love numbers, 62
  marine tide, 57, 62
  phase lag, 63
  semidiurnal tide, 58, 59
  solar tide, 59
  spring and neap tides, 61
  synchroneity of rotations, 12, 64
time units
  convention for ages and dates, 242
  day, sidereal and solar, 235
  month, sidereal and solar, 236
  second, ephemeris time, 236
  universal time, 236
  year, sidereal and tropical, 236
timescale
  biostratigraphical, 236

  - chronostratigraphical, 237
  - geological, 236, 237
    - eons, eras, periods, and epochs, 237
  - geomagnetic polarity, 68, 385, 387, 389
- Titan, moon of Saturn, 16
- titanomagnetite, 358, 359, 387
  - oceanic basalt magnetization, 359
- transform faults, 34, 199, 200
  - fracture zones, 34
  - location of Euler rotation pole, 34
  - New Zealand Alpine fault, 34
  - oblique slip, 200
  - relative motion, 201
  - seismicity, 34
- travel time
  - dependence on epicentral distance, 220
- trench, deep-sea, 26, 31, 33, 36
  - geoid undulation, 78
  - gravity anomaly, 116
  - heat flow, 279
  - lithosphere flexure, 139
  - role in subduction, 44
  - Vening Meinesz isostasy model, 120
- triple junctions, 35
  - evolution of northeast Pacific plates, 37
  - formation of San Andreas fault, 38
  - interplate velocity diagram, 35
  - possible types, 37
  - relative velocity diagram, 37
  - stability criteria, 36, 37
- true polar wander, 286
- tsunami, 214
  - early warning system, 215
  - Pacific Ocean earthquakes, 214, 215
  - period of wave, 213

  Sumatra earthquake, 214
  Tohoku earthquake, 215
  velocity dependence on ocean depth, 213, 214
  wavelength, 214

Uranus, 17, 339
  composition, 17
  discovery, 9
  internal structure, 17
  magnetic field, 17, 339
  obliquity of rotation axis, 10, 17
  rings, 17
  spin axis obliquity, 339

Van Allen radiation belts, 323
Venus, 11, 337
  magnetosphere, 337
  retrograde rotation, 11
  surface temperature, 11
  Venus Express spacecraft, 11
very long baseline interferometry (VLBI), 66, 76, 83
Vine–Matthews–Morley hypothesis, 29, 32, 387
virtual geomagnetic pole, 23, 356
viscoelasticity, 135, 142
viscosity, 132, 134
  kinematic, 281
  mantle, 140, 140, 143
  mantle, lower, 142
  mantle, upper, 141
viscous flow, 134, 135
  Newtonian fluid, 132, 134, 140
viscous flow, in mantle, 21, 43
volcanic islands
  arcuate chains at plate margins, 38
  linear intraplate chains, 40
volcanic tremor, 192

Wadati diagram, 194
Wadati–Benioff seismic zone, 33, 196, 201, 231
wave equation
  D’Alembert’s principle, 151
  numerical solutions, 173
    finite-difference method, 174, 175
    finite-element analysis, 175
  one-dimensional
    compressional wave, 149
    shear wave, 150
  properties of waves, 151
  three-dimensional equations, 394
    compressional wave, 395
    shear wave, 395
waveform tomography, 224, 225
wavelength filtering, 105

Young’s modulus, 128, 129, 131
Yungay landslide, 213

Zeeman effect, 334, 343